Leveraging Semi-Supervised Learning for Multimodal Hate Speech Data Annotation and Detection

Rathi Adarshi Rammohan[1], Zhao Ren[1], Dominik Puchała[2], Aleksandra Świderska[2], Dennis Küster[1], Tanja Schultz[1]

[1]Cognitive Systems Lab, University of Bremen, Germany
[2]Faculty of Psychology, University of Warsaw, Poland
rammohan@uni-bremen.de, zren@uni-bremen.de, dominik.puchala@psych.uw.edu.pl,
aleksandra.swiderska@psych.uw.edu.pl, dennis.kuester@uni-bremen.de, tanja.schultz@uni-bremen.de

Abstract

While the Internet and social media have fundamentally transformed our lives, they can also rapidly spread hate speech, i.e., derogatory statements targeting individuals or groups based on their immutable characteristics. Automatic detection systems could help limit this harmful phenomenon. However, the lack of large-scale annotated datasets remains a major bottleneck for developing better algorithms. In this work, we employ Semi-Supervised Learning (SSL) to leverage the advantages of limited labeled data alongside large amounts of unlabeled data. We apply three SSL approaches, Fix-match, Full-match, and All-match learning, to enhance the performance of end-to-end pre-trained speech and text models for hate speech detection. Our findings indicate that SSL methods enhance the performance, achieving F1 scores of 0.851 on speech, 0.957 on text, and 0.959 with multimodal fusion. Furthermore, we analyze the impact of different weak augmentation strategies on labeled data and assess the quality of generated pseudo-labels to evaluate their potential use in data annotation.

Keywords: Semi-Supervised Learning, Hate Speech Detection, Pseudo-Labeling, Data Annotation

1. Introduction

When considering how much the Internet, social media, and infotainment have become part of modern daily life (Sparrow and Chatman, 2013; Wirz and Zai, 2025), encounters with heated emotional discussions (Garcia et al., 2016) or hateful contents (Dreißigacker et al., 2024) may seem almost unavoidable. In Europe, network providers and regulatory bodies have established guidelines to maintain the quality of content on their platforms and to remove any material that does not comply with these standards (European Commission, 2024). One of the most common violations of EU regulations reported on the Digital Services Act (DSA) Transparency Database[1] is hate speech, which is defined as a form of verbal aggression that is capable of inciting or propagating hatred, discrimination, or violence against individuals or groups based on certain characteristics such as race, sexual orientation, or nationality. The proliferation of hate speech can have severe consequences among both the majority group and those targeted. For example, exposure to hate speech among the majority group desensitizes people to further hateful content, which, in turn, results in greater prejudice against victims of hate speech (Soral et al., 2018, 2023). Exposure to hate speech among minorities is associated with a deterioration in their well-being (Zochniak et al., 2023; Wypych and Bilewicz, 2024).

With the large amounts of data available on online platforms, it is resource-intensive and mentally exhausting to manually identify hate speech. Automatic content moderation algorithms may help to automate and improve this process (Drolsbach and Pröllochs, 2024). However, one of the key challenges faced by researchers working on hate speech detection is the limited availability of annotated databases, which hinders the development of accurate systems that effectively detect hate speech while minimizing false alarms.

A commonly adopted approach to improve model performance is to use Semi-Supervised Learning on a small amount of labeled data along with large-scale unlabeled data (SSL) (Chapelle et al., 2009). In this method, a model that has been trained on the labeled data generates pseudo-labels on the unlabeled data, which is then used to train the model for classification or clustering tasks. Various types of SSL techniques are available. For example, pseudo-labels are generated and models are trained iteratively, or a portion of the unlabeled data is chosen based on a confidence threshold on their pseudo-labels and then jointly used with labeled data for supervised learning (Reddy et al., 2018; Ouali et al., 2020). In the speech processing domain, SSL has been applied to speech foundation models to perform tasks such as automatic speech recognition (Synnaeve et al., 2019; Park and Hain, 2025) and speech emotion recognition (Ren et al., 2025; Huang et al., 2018; Zhang et al., 2021). Inspired by the work in (Ren et al., 2025), we leverage Fix-match (Sohn et al.,

[1]https://transparency.dsa.ec.europa.eu/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 10266–10275
11–16 May 2026.

2020), Full-match (Chen et al., 2023) learning techniques, and additionally employ All-match (Wu and Cui, 2024) learning to perform hate speech detection.

The main objectives of this contribution are as follows. First, we aim to apply and evaluate the effectiveness of the three Semi-Supervised Learning techniques, Fix-match, Full-match, and All-match, for speech- and text-based hate speech detection using a combination of labeled and large-scale unlabeled Polish data. Second, we compare the impact of different data augmentation methods applied to both labeled and unlabeled datasets. Finally, we validate the quality of a subset of the generated pseudo-labels to examine whether these techniques can be reliably used for data annotation.

2. Related Works

SSL approaches have been applied in hate speech detection research, predominantly focusing on textual content. For example, in the works of (Cahyana et al., 2022) and (Saifullah et al., 2024), the detection and annotation of hate speech were performed on YouTube comments. Natural language processing techniques were used to extract linguistic features and train machine learning models on labeled data. Pseudo-labels were then generated for the unlabeled data based on a confidence threshold applied to the predicted hate scores. The entire process was conducted iteratively and incorporated subseqent manual annotation of samples that did not meet the threshold criterion.

Similarly, (Ludwig et al., 2024) employed another strategy in addition to the threshold-based method with a ratio-based approach, which considered the percentage of samples belonging to the target (i.e., hate) class when generating pseudo-labels. In (Tung et al., 2023), multimodal memes, comprising text and images, were analyzed. Using a cross-modal autoencoder, representations were extracted from the labeled data and subsequently fused with those derived from the unlabeled samples. In the work of (Das et al., 2025), both Generative Adversarial Networks (GAN) and Fix-match were applied as SSL methods for the detection of hate speech in tweets. Here, GANs were used to generate synthetic tweets for data augmentation, and their integration with Fix-match led to optimal performance.

In contrast, our work explores hate speech detection in speech and text modalities using three SSL methods, Fix-match, Full-match, and All-match, to assess their effectiveness in leveraging unlabeled data for multimodal hate speech analysis.

3. Methodology

3.1. Speech and Text Foundation Models

Since the introduction of transformer models, conventional feature-based methods have been replaced by end-to-end pre-trained acoustic and linguistic models for speech processing tasks such as speech recognition, speaker verification, and speech emotion recognition (Arora et al., 2024; Phukan et al., 2023). Using a self-supervised learning approach, these models are trained on large datasets and are able to extract contextual and semantic information from raw speech and text data. In the present work, we employed Wav2Vec2.0-XLSR (Conneau et al., 2020) (large[2]) and XLM-RoBERTa (Conneau et al., 2019) (base[3]) models, both trained on multilingual data for processing our speech and text inputs, respectively. Additionally, we added a linear classifier layer to perform hate speech detection.

3.2. Semi-Supervised Learning

Semi-Supervised Learning aims to train a model f based on the labeled data $D_l : \{X_l, y_l\}$ and the unlabeled data $D_u : \{X_u\}$. The model is trained on D_l in a supervised learning manner and is able to automatically annotate D_u to enhance its ability to learn representations gained from more data samples. In this section, we discuss and compare the following three Semi-Supervised Learning approaches, including Fix-match learning, Full-match learning, and All-match learning. Figure 1 represents the overall pipeline of the three Semi-Supervised Learning methods.

In the three Semi-Supervised Learning approaches, the labeled and unlabeled data are first augmented with a weak augmentation process to increase data size and variability, thereby enhancing model robustness. The augmented labeled and unlabeled data are represented as $D_l^w : \{X_l^w, y_l\}$ and $D_u^w : \{X_u^w\}$, respectively. In addition, the unlabeled data is strongly augmented into $D_u^s : \{X_u^s\}$. Herein, strong augmentation leads to a bigger data difference before and after augmentation, e. g., adding noise to revise every speech frame, compared to weak augmentation.

3.2.1. Fix-match Learning

In addition to training a model f on the weakly-augmented labeled data in a supervised learning manner, Fix-match learning (Sohn et al., 2020)

[2]https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[3]https://huggingface.co/FacebookAI/xlm-roberta-base

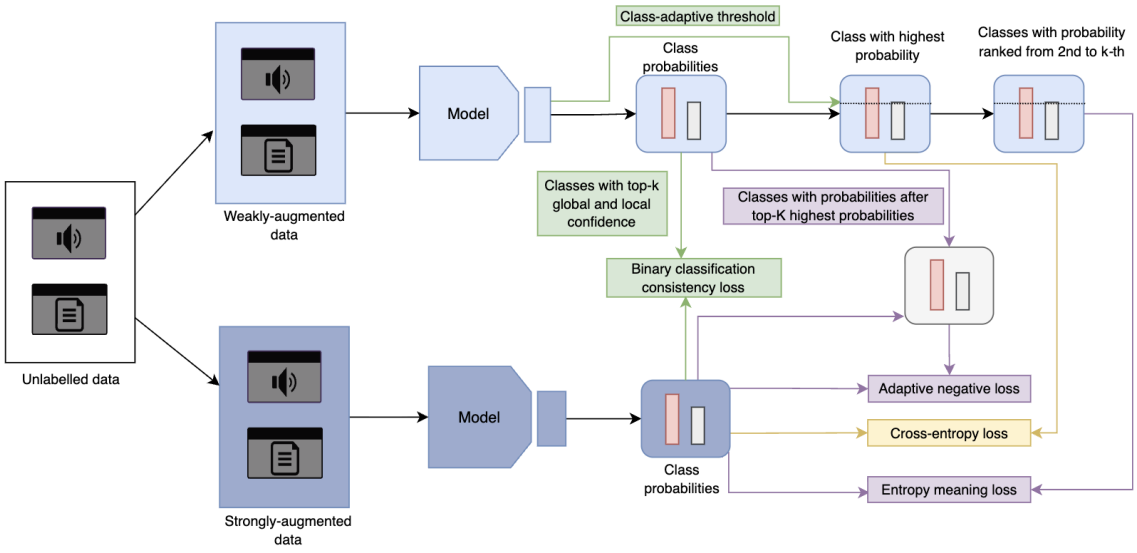

Figure 1: Diagrammatic representation of the Semi-Supervised Learning methods employed in the present work. The yellow, violet and green lines represent the loss functions used in Fix-match learning, Full-match and All-match learning, respectively.

aims to train f on the strongly-augmented unlabeled data by generating pseudo labels. For each unlabeled speech sample, the pseudo label is generated when the highest class probability predicted from the weakly-augmented sample is larger than a threshold ϵ. The loss function in Fix-match learning is defined by

$$\mathcal{L}_{\text{fix}} = \mathcal{L}_{\text{ce}}(X_l^w, y_l) + \lambda_{\text{fix}} \mathbb{1}(\max f(X_u^w) > \epsilon) \mathcal{L}_{\text{ce}}(X_u^s, f(X_u^w)), \quad (1)$$

where $\mathcal{L}_{\text{ce}}$ is the cross-entropy loss, and λ_{fix} is a constant factor to balance the two cross-entropy loss functions.

3.2.2. Full-match Learning

Fix-match learning may annotate only a portion of unlabeled samples, since the highest class probabilities have to pass a threshold. The unlabeled samples are not used during training when $\max f(X_u^w) \leq \epsilon$. To leverage all of the unlabeled samples, Full-match learning (Chen et al., 2023) involves two additional loss functions, namely adaptive negative loss and entropy meaning loss. The top-k accuracy of $f(X_u^s)$ is first calculated when $f(X_u^w)$ is considered as the pseudo labels. The value of k is then determined when the top-k accuracy is larger than a threshold of τ. Given the top-k accuracy, the class probabilities of each weakly-augmented unlabeled sample are ranked decreasingly.

Adaptive negative loss. The adaptive negative loss aims to minimize the probabilities of the classes ranked after k, thereby increasing the model confidence on the top-k classes. The adaptive negative loss is calculated by

$$\mathcal{L}_{\text{adp}} = -\sum_{c=1}^{C} \log(1 - f(X_u^s)) \mathbb{1}(\text{R}(f(X_u^w)) > k), \quad (2)$$

where C is the number of classes, and R is the rank of the probabilities.

Entropy meaning loss. For each unlabeled sample, the class probabilities ranked from 2 to k are regulated to be as low as possible compared to the top-1 probability. Meanwhile, the 2^{th}-k^{th} class probabilities are trained to share similar values to improve the confidence of the model in the top-1 class. For such a purpose, the entropy meaning loss is designed as

$$\mathcal{L}_{\text{ent}} = -\sum_{c=1}^{C} \mathbb{1}(\text{R}(f(X_u^w)) \in [2, k])(y_u^e(\log f(X_u^s) + (1 - y_u^e) \log(1 - f(X_u^s)))), \quad (3)$$

$$y_u^e = \frac{1 - \sum_{c=1}^{C} \mathbb{1}(\text{R}(f(X_u^w)) \in [2, k]) f(X_u^s)}{k - 1}. \quad (4)$$

The overall Full-match loss is calculated by

$$\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{fix}} + \lambda_{\text{adp}} \mathcal{L}_{\text{adp}} + \lambda_{\text{ent}} \mathcal{L}_{\text{ent}}, \quad (5)$$

where λ_{adp} and λ_{ent} are constant factors.

3.2.3. All-match Learning

Compared to previous methods, in All-match learning (Wu and Cui, 2024), the confidence threshold is dynamically determined by the model's predictions and the classifier weights, referred to as the *Class-Adaptive Threshold (CAT)*. This is achieved in two steps:

1) Global estimation: This step assesses the overall learning status of the model. The mean confidence of the predictions on the weakly augmented unlabeled samples is used as an estimate, denoted as τ_t, and is updated during training using an exponential moving average (EMA) with a momentum factor m_τ:

$$\tau_t = m_\tau \tau_{t-1} + (1 - m_\tau) \frac{1}{B} \sum_{i=1}^{B} \max f(X_{u,i}^w), \quad (6)$$

where $f(X_{u,i}^w)$ represents the model's prediction probabilities for the weakly augmented unlabeled samples. We initialize $\tau_0 = 1/C$, where C is the number of classes.

2) Class-specific (local) estimation: In this step, the threshold is adjusted to account for underfitting classes. The L2 norm of the classifier weights is combined with the global estimate to obtain the overall class-adaptive threshold:

$$\tau_t(c) = \tau_t \cdot \frac{\|W_c\|_2}{\max_k \|W_k\|_2}, \quad (7)$$

where W_c denotes the weight vector of class c and and k indexes over all C classes.

Once the class-adaptive thresholds are determined, pseudo-labels are generated for the unlabeled samples, and the unlabeled loss $\mathcal{L}_\mathsf{u}$ is computed only for those samples whose prediction confidence exceeds the corresponding $\tau_t(c)$. For all unlabeled data, an additional step called *Binary Classification Consistency (BCC)* regularization is introduced to maintain consistency between the weakly and strongly augmented predictions. For each unlabeled sample, the top-K confidence classes are treated as candidate (positive) classes, while the remaining ones are considered negative. For both weak and strong augmentations, the binary distributions over the candidate and negative classes are computed as:

$$b_i^{(w)} = \Big[\sum_{c \in C_i} f_c(x_i^{(w)}),\ 1 - \sum_{c \in C_i} f_c(x_i^{(w)}) \Big], \quad (8)$$

$$b_i^{(s)} = \Big[\sum_{c \in C_i} f_c(x_i^{(s)}),\ 1 - \sum_{c \in C_i} f_c(x_i^{(s)}) \Big]. \quad (9)$$

The BCC regularization loss is then defined as:

$$\mathcal{L}_\mathrm{b} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{t \in \{0,1\}} b_{i,t}^{(w)} \log b_{i,t}^{(s)}. \quad (10)$$

Finally, the overall All-match loss is computed as:

$$\mathcal{L}_\mathsf{all} = \mathcal{L}_\mathsf{ce} + \lambda_\mathsf{u} \mathcal{L}_\mathsf{u} + \lambda_\mathsf{b} \mathcal{L}_\mathsf{b}. \quad (11)$$

3.2.4. Data Augmentation

Following the approach in our prior study (Ren et al., 2025), we applied two levels of augmentation, weak and strong, on both speech and text samples, and compared the performance of different weak augmentation methods. For the speech data, we used two weak augmentations: *Flipping*, where a randomly chosen segment (up to 6.25 s) was reversed, and *Time masking*, where a random segment (up to 30 k frames) was set to zero. For strong augmentation, we applied *Gaussian noise* with a scaling factor of 0.05.

For the text data, we used two weak augmentations: *Swapping*, where adjacent words were randomly interchanged, and *Deleting*, where random partial words were removed. For strong augmentation, we applied *Contextual substitution*, where the top n semantically similar words were substituted using contextual word embeddings derived from XLM-RoBERTa (base).

4. Experiments and Results

4.1. Dataset

For this task, we used two sets of data collected from similar sources[4]. The fully annotated Warsaw Multimodal Hate Speech (WMHS) Database served as our labeled dataset, which was collected from social media platforms such as Facebook, YouTube, Tiktok and Banbye[5] through targeted keyword searches. The database comprises 770 samples (approximately 9 hours) of Polish hate and 763 samples (approximately 9 hours) of non-hate speech, and corresponding transcriptions, targeting various social groups. All samples were annotated by human experts (Puchała et al., 2025).

To obtain unlabeled data, we used a similar set of keywords and scraped content from Banbye over a period of three months (September to November) 2023, gathering around 6900 videos. For this work, we considered only videos with a maximum duration of 15 minutes. The audio was extracted using Moviepy[6] python package and transcribed into text using OpenAI's Whisper 'large' model (Radford et al., 2023). After removing English and other non-Polish content, this process yielded 778 samples,

[4]Both the labeled and unlabeled datasets are available upon request.

[5]Banbye is a Polish social media platform that does not remove content containing hate speech.

[6]https://pypi.org/project/moviepy/

i.e., approximately 78 hours of unlabeled speech and text data.

4.2. Model Training and Evaluation

The data was split into training, validation, and test sets with a ratio of 70:15:15, ensuring an equal class distribution across all three subsets. Both the speech and text models were trained using the Adam optimizer with a learning rate of 3×10^{-5} and Cross-Entropy Loss for 5 epochs. A batch size of 8 was used for all speech models, except for All-match, where the batch size was 4, and 16 for the text models, depending on memory requirements.

The raw speech signals were segmented into 5-second clips with an overlap of 1 second, and majority voting was applied to obtain the final prediction for each complete speech sample. The performance of the models was evaluated using the F1 score.

The loss function coefficients in Equations 1 and 5 were set as $\lambda_{\text{fix}} = \lambda_{\text{adp}} = \lambda_{\text{ent}} = 0.5$ and in Equation 11, the constant factors were set as $\lambda_{\text{u}} = \lambda_{\text{b}} = 1$. The hyperparameters were configured as $\epsilon = 0.95$, $\sigma = 0.99$ and $m_\tau = 0.99$.

4.3. Results and Discussion

4.3.1. Performance of Speech and Text Models

Table 1 presents the performance of the speech and text models for hate speech detection across the different SSL approaches. The baseline depicts the performance of the pre-trained Wav2Vec2.0-XLSR and the XLM-RoBERTa models fine-tuned on the WMHS data. The three SSL techniques were subsequently applied to utilize both labeled and unlabeled data, and the results were compared.

Overall, the text model outperformed the speech models in detecting hate speech. For the baseline, the speech model achieved an F1 score of 0.834, and the text model yielded 0.922, both above the chance level (50%). Hate speech detection relies heavily on the semantic context to capture the offensiveness and the target of the hate speech (Mathew et al., 2021). However, the speech model's performance also shows that the acoustic properties alone already support a fairly robust distinction between hate and non-hate speech. This suggests that speech may contribute to distinguishing between types or shades of hate speech.

4.3.2. Semi-Supervised Learning for Hate Speech Detection

The application of SSL techniques further improved the performance of the models. Across both modalities, Full-match and All-match produced higher F1 scores, demonstrating that the use of adaptive learning and utilizing all the unlabeled samples enhances the effectiveness of the classification. For the speech modality, All-match yielded the highest F1 score of 0.851, while for the text modality, Full-match obtained an F1 score of 0.957.

Based on the F1 scores, we chose the best-performing Fix-match, Full-match and All-match models, respectively for speech and text. We then applied decision-level fusion, wherein the average of the speech and text probabilities was computed to obtain the final predictions for each sample. Additionally, the overall best-performing speech and text models were also fused, indicated as 'Best 2' in Table 1. The fused Full-match achieved an F1 score of 0.959, outperforming the unimodal results and highlighting the benefits of leveraging both modalities for hate speech detection.

In (Wu and Cui, 2024), All-match was shown to outperform all the other SSL methods, which was attributed to the effective use of all the samples by applying class-adaptive threshold and binary classification consistency regularization on the unlabeled samples. The speech models in the present work demonstrate a similar trend. However, for the text modality and in fusion of modalities, Full-match performed better. Since the unlabeled data in our study were scraped from social media platforms, their source diversity and quality variations may have impacted the effectiveness of methods that incorporate all samples.

4.3.3. Ablation study

We conducted a systematic ablation study to assess the influence of data augmentation on the labeled data. During training, in addition to applying weak and strong augmentations to the unlabeled data, we also applied weak augmentation to the labeled data.

As shown in Table 1, we observe that, in general, training without augmentation yields better results compared to training with augmentation. Specifically, for both the speech and text modalities, the Full-match approach consistently achieved higher or equal performance without augmentation. This suggests that augmenting the labeled data may introduce variability that drifts from the natural data distribution, which can negatively affect performance, especially when the distinctions between classes are subtle or implicit.

However, when the modalities were fused, models trained with augmentation exhibited improved

Table 1: The performance of the speech (Wav2Vec2.0-XLSR) and text (XLM-RoBERTa) models in terms of F1 score across the different Semi-Supervised Learning techniques and augmentation strategies.

Model	Method	Augment	without augmentation		with augmentation	
			Valid	Test	Valid	Test
Wav2Vec2.0-XLSR	Baseline	—	0.846	0.834	0.846	0.834
	Fix-match	Flipping	0.843	0.812	0.855	**0.848**
	Fix-match	Time masking	0.832	0.817	0.850	0.808
	Full-match	Flipping	0.834	0.838	0.812	0.797
	Full-match	Time masking	0.834	0.839	0.838	0.834
	All-match	Flipping	0.808	0.835	0.838	0.839
	All-match	Time masking	0.842	**0.851**	0.825	0.813
XLM-RoBERTa	Baseline	—	0.912	0.922	0.912	0.922
	Fix-match	Swapping	0.921	0.926	0.908	0.943
	Fix-match	Deleting	0.917	**0.957**	0.908	0.882
	Full-match	Swapping	0.917	0.935	0.908	0.886
	Full-match	Deleting	0.921	**0.957**	0.926	**0.957**
	All-match	Swapping	0.913	0.939	0.900	0.935
	All-match	Deleting	0.913	0.930	0.913	0.952
Fusion	Fix-match	—	0.920	**0.952**	0.929	0.948
	Full-match	—	0.938	0.908	0.938	**0.959**
	All-match	—	0.947	0.926	0.929	0.956
	Best 2	—	0.942	0.939	0.947	0.956

performance, indicating that the increased data complexity helped the fusion model better generalize across modalities.

Additionally, we compared two different weak augmentation techniques. Overall, time masking and deleting produced better performances for the speech and text modalities, respectively. This suggests that information removal, rather than distortion, can benefit model learning by encouraging robustness to missing or occluded features.

4.3.4. Validation of Pseudo-Labels

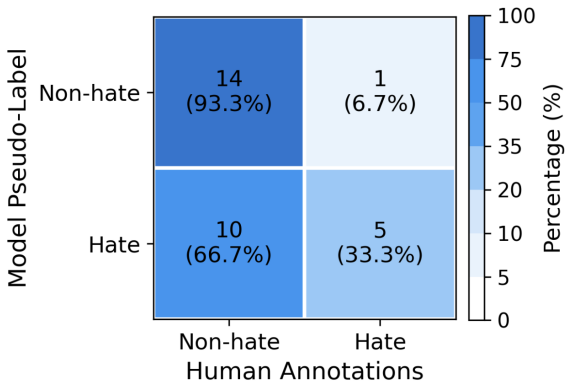

Figure 2: Confusion matrix of pseudo-labels against the human annotations.

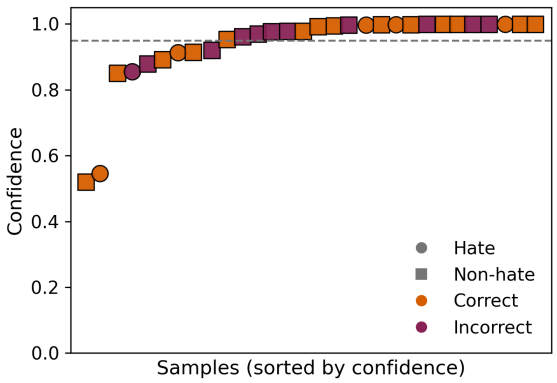

Figure 3: Scatter plot of the confidence scores of the generated pseudo-labels. Correct and incorrect predictions are shown in green and red, respectively. The dotted horizontal line indicates the threshold applied during training.

The text model with Full-match SSL approach produced the best results across both conditions - with and without augmentation of labeled data. To reduce data variability, we consider the model trained without augmentation. Using this model, we generated the pseudo-labels for the unlabeled dataset along with their corresponding confidence scores. Out of the total 778 unlabeled samples, 30 samples (15 hate and 15 non-hate) with varying confidence scores were randomly selected. These pseudo-labels were verified by a human expert, who annotated the WMHS data. Figures 2 and 3

presents the confusion matrix of the pseudo-labels and the human validated labels and the scatter plot of the confidence scores of the pseudo-labels. Among the hate samples, 5 out of 15 were correctly labeled, and for the non-hate samples, almost all were correctly labeled, except one sample. Most of the hate speech samples labeled by the model as hate speech were identified by the human as non-hate speech. These findings imply that the models are more effective in detecting non-hate speech than hate speech, potentially due to the sometimes subtle linguistic differences separating hate speech from non-hate speech.

While the SSL methods effectively improved model performance using unlabeled data, they also raise concerns regarding the reliability of pseudo-labels for direct use in data annotation. Some recent research suggests that hate speech has increasingly evolved into more implicit and subtle forms, often avoiding explicit offensive terms (ElSherief et al., 2021; Lin, 2022; Mohammed, 2025). Detecting such more subtle speech requires understanding the context and correctly identifying the target social group. Furthermore, it can be argued that the cost of misclassifying hate speech as non-hate is higher than the reverse. Thus, when non-hate speech is mistakenly classified as hate speech by a content moderation algorithm, the author typically has an opportunity to appeal and request review (Casarosa, 2023). In comparison, when hate speech is missed by an algorithm, recourse to human moderation means that, at minimum, the human who flags a statement as hate speech has already been exposed to it.

These findings suggest that SSL-based models could serve as a foundation for generating preliminary annotations, which can then be refined through an iterative human-in-the-loop annotation process to improve the accuracy and reliability of hate speech identification.

5. Conclusion and Future Work

Hate speech remains a serious social issue, in particular for democratic societies that value human- and minority rights as well as individual freedom of expression. The ethical development of well-balanced automatic hate speech detection algorithms could therefore provide an essential building block of our answer to the rapid spread of hate speech via social media. In this work, we leveraged the annotated Polish WMHS database and unlabeled data from Banbye to apply three SSL methods for hate speech detection using pre-trained speech and text foundation models. Our results show that SSL methods can successfully leverage unlabeled data to improve model performance, with the Full-match method on fused text- and speech models yielding the overall best results.

As demonstrated by the ablation study, training without augmentation generally produced better results for hate speech detection, suggesting that it may be important to preserve the original data distribution. Additionally, we analyzed a subset of the pseudo-labels generated by the text model using Full-match and observed that these models were more effective in detecting non-hate speech than hate speech. This suggests that our approach may be conservative with respect to avoiding mistakes where harmless material is falsely labeled as hate-speech, thus erring on the side of rather allowing a piece of hate speech to pass than to infringe upon expressions that might still be harmless.

In conclusion, SSL models demonstrate strong potential for integration into a human-in-the-loop annotation workflow. Such a workflow could facilitate efficient annotation of larger hate speech data corpora and thus more robust hate speech detection. Future work could further aim to improve the detection of subtle and implicit hate speech. Here, e.g., large language models (LLMs) could be explored for semi-automatic data annotation and enhanced with semantic context knowledge through adaptive learning.

6. Ethical Statement

Hate speech has serious implications for both individuals and society(Soral et al., 2018; Zochniak et al., 2023). With the widespread accessibility of the Internet, users are increasingly vulnerable to exposure to harmful content. Research on detecting and understanding hate speech is essential to curbing its spread and improving online safety.

The dataset used for our research consists of videos collected from various sources. The dataset lacks metadata, but some videos include identifiable individuals. However, our focus was on extracting the acoustic characteristics of speech rather than speaker identity for hate speech detection.

Exposure to hateful content, whether through annotation, analysis, or model training, poses a mental health risk for researchers. We advise researchers and annotators to engage with the content in controlled environments, taking regular breaks and seeking support if needed.

There is a risk that hate speech detection models, including ours, could be misused to identify and amplify hateful content targeting specific minority groups. While our intent is purely academic and socially beneficial, future misuse with malicious intentions is possible. We strongly advocate that this technology be used responsibly and only in research contexts that uphold ethical safeguards. To mitigate the risk, annotated hate speech databases, such as the Warsaw Multimodal

Hate Speech (WMHS) Database (Puchała et al., 2025), are often made available only upon request by researchers.

Our work represents a step toward efficiently detecting hate speech in speech and text data on social media platforms. We recognize the ethical complexities inherent in this domain and encourage open discourse, responsible data handling, and cautious application to mitigate potential harms.

7. Acknowledgments

This research was funded by Deutsche Forschungsgemeinschaft (DFG; project number 465129985) and Polish National Science Centre (Narodowe Centrum Nauki, NCN; grant number 2020/39/G/HS6/00231), under the Beethoven CLASSIC 4 Polish-German funding initiative.

8. Bibliographical References

Siddhant Arora, Ankita Pasad, Chung-Ming Chien, Jionghao Han, Roshan Sharma, Jee-weon Jung, Hira Dhamyal, William Chen, Suwon Shon, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2024. On the Evaluation of Speech Foundation Models for Spoken Language Understanding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11923–11938, Bangkok, Thailand. Association for Computational Linguistics.

Nur Heri Cahyana, Shoffan Saifullah, Yuli Fauziah, Agus Sasmito Aribowo, and Rafal Drezewski. 2022. Semi-supervised Text Annotation for Hate Speech Detection Using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency. *International Journal of Advanced Computer Science and Applications*, 13(10).

Federica Casarosa. 2023. Out-of-court Dispute Settlement Mechanisms for Failures in Content Moderation. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, 14:391.

O. Chapelle, B. Scholkopf, and A. Zien, Eds. 2009. Semi-Supervised Learning. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Yuhao Chen, Xin Tan, Borui Zhao, Zhaowei Chen, Renjie Song, Jiajun Liang, and Xuequan Lu. 2023. Boosting Semi-Supervised Learning by Exploiting All Unlabeled Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7548–7557.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. *arXiv preprint arXiv:2006.13979*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116*.

Tithi Rani Das, Tanjim Mahmud, Abubokor Hanip, and Mohammad Shahadat Hossain. 2025. Harmful Tweet Detection using Supervised and Semi-Supervised Learning Techniques. In *2025 International Conference on Inventive Computation Technologies (ICICT)*, pages 1421–1427.

Arne Dreißigacker, Philipp Müller, Anna Isenhardt, and Jonas Schemmel. 2024. Online Hate Speech Victimization: Consequences for Victims' Feelings of Insecurity. *Crime Science*, 13(1):4.

Chiara Patricia Drolsbach and Nicolas Pröllochs. 2024. Content Moderation on Social Media in the EU: Insights from the DSA Transparency Database. In *Proceedings of the ACM Web Conference 2024*, pages 939–942.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. *arXiv preprint arXiv:2109.05322*.

European Commission. 2024. The Digital Services Act (DSA). Official website of the European Commission. Accessed: 2025-10-14.

David Garcia, Arvid Kappas, Dennis Küster, and Frank Schweitzer. 2016. The Dynamics of Emotions in Online Interaction. *Royal Society Open Science*, 3(8):160059.

Jian Huang, Ya Li, Jianhua Tao, Zheng Lian, Mingyue Niu, and Jiangyan Yi. 2018. Speech Emotion Recognition Using Semi-Supervised Learning with Ladder Networks. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–5. IEEE.

Jessica Lin. 2022. Leveraging World Knowledge in Implicit Hate Speech Detection. *arXiv preprint arXiv:2212.14100*.

Florian Ludwig, Klara Dolos, Ana Alves-Pinto, and Torsten Zesch. 2024. Unraveling the dynamics of semi-supervised hate speech detection:

The impact of unlabeled data characteristics and pseudo-labeling strategies. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1974–1986, St. Julian's, Malta. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Dima Mohammed. 2025. Where the hate lies in soft hate speech: the argumentative potential in hostile public spheres. *Critical Discourse Studies*, pages 1–14.

Yassine Ouali, Céline Hudelot, and Myriam Tami. 2020. An Overview of Deep Semi-Supervised Learning. *arXiv preprint arXiv:2006.05278*.

Chanho Park and Thomas Hain. 2025. Semi-Supervised Learning for Automatic Speech Recognition with Word Error Rate Estimation and Targeted Domain Data Selection. In *Proceedings of Interspeech 2025*. International Speech Communication Association (ISCA).

Orchid Chetia Phukan, Arun Balaji Buduru, and Rajesh Sharma. 2023. A Comparative Study of Pre-trained Speech and Audio Embeddings for Speech Emotion Recognition. *arXiv preprint arXiv:2304.11472*.

Dominik Puchała, Rathi Adarshi Rammohan, Dennis Küster, Tanja Schultz, and Aleksandra Świderska. 2025. The Warsaw Multimodal Hate Speech Database (WMHS): Development, initial validation, and baseline automatic detection. Manuscript submitted for publication.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of International Conference on Machine Learning*. 27 pages.

YCAP Reddy, P Viswanath, and B Eswara Reddy. 2018. Semi-Supervised Learning: A Brief Review. *International Journal of Engineering & Technology*, 7(1.8):81.

Zhao Ren, Rathi Adarshi Rammohan, Kevin Scheck, Sheng Li, and Tanja Schultz. 2025. End-to-end Acoustic-linguistic Emotion and Intent Recognition Enhanced by Semi-Supervised Learning. *arXiv preprint arXiv:2507.07806*.

Shoffan Saifullah, Rafał Dreżewski, Felix Andika Dwiyanto, Agus Sasmito Aribowo, Yuli Fauziah, and Nur Heri Cahyana. 2024. Automated Text Annotation Using a Semi-supervised Approach with Meta Vectorizer and Machine Learning Algorithms for Hate Speech Detection. *Applied Sciences*, 14(3):1078.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. 2020. Fixmatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems*, 33:596–608.

Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. 2018. Exposure to Hate Speech Increases Prejudice Through Desensitization. *Aggressive Behavior*, 44(2):136–146.

Wiktor Soral, Aleksandra Świderska, Dominik Puchała, and Michał Bilewicz. 2023. Desensitization to Hate Speech: Examination Using Heart Rate Measurement. *Aggressive Behavior*, 50(1):e22118.

Betsy Sparrow and Ljubica Chatman. 2013. Social Cognition in the Internet Age: Same As It Ever Was? *Psychological Inquiry*, 24(4):273–292. Publisher: Routledge _eprint: https://doi.org/10.1080/1047840X.2013.827079.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end ASR: From Supervised to Semi-Supervised Learning with Modern Architectures. *arXiv preprint arXiv:1911.08460*.

Pham Thai Hoang Tung, Nguyen Tan Viet, Ngo Tien Anh, and Phan Duy Hung. 2023. SemiMemes: A Semi-Supervised Learning Approach for Multimodal Memes Analysis. In *International Conference on Computational Collective Intelligence*, pages 565–577. Springer.

Dominique S. Wirz and Florin Zai. 2025. Infotainment on Social Media: How News Companies Combine Information and Entertainment in News Stories on Instagram and TikTok. *Digital Journalism*, 13(7):1249–1270. Publisher: Routledge _eprint: https://doi.org/10.1080/21670811.2025.2464062.

Zhiyu Wu and Jinshi Cui. 2024. AllMatch: Exploiting All Unlabeled Data for Semi-Supervised Learning. *arXiv preprint arXiv:2406.15763*.

Michał Wypych and Michał Bilewicz. 2024. Psychological Toll of Hate Speech: The Role of Acculturation Stress in the Effects of Exposure to Ethnic Slurs on Mental Health among Ukrainian Immigrants in Poland. *Cultural Diversity & Ethnic Minority Psychology*, 30(1):35–44.

Sheng Zhang, Min Chen, Jincai Chen, Yuan-Fang Li, Yiling Wu, Minglei Li, and Chuanbo Zhu. 2021. Combining Cross-Modal Knowledge Transfer and Semi-Supervised Learning for Speech Emotion Recognition. *Knowledge-Based Systems*, 229:107340.

Kamila Zochniak, Oliwia Lewicka, Zuzanna Wybrańska, and Michał Bilewicz. 2023. Homophobic Hate Speech Affects Well-Being of Highly Identified LGBT People. *Journal of Language and Social Psychology*, 42(4).