MuLTa-Telegram: A Fine-Grained Italian and Polish Dataset for Hate Speech and Target Detection

Elisa Leonardelli[1,*], Camilla Casula[1], Sebastiano Vecellio Salto[1], Joanna Ewa Bak[1], Elisa Muratore[1], Anna Kolos[2], Thomas Louf[1] and Sara Tonelli[1]

[1]Fondazione Bruno Kessler (FBK), Via Sommarive 18, 38123 Trento, Italy
[2]NASK National Research Institute, ul. Kolska 12, 01-045 Warsaw, Poland

Abstract

This paper introduces the *MuLTa-Telegram* dataset, a *Mu*lti- *L*ingual and multi-*Ta*rget dataset specifically developed to detect hate speech on *Telegram*, an understudied yet influential platform in which extremist and fringe content can be found. The dataset contains about 4,000 Telegram messages in Italian and Polish, annotated for the presence of hate speech and its targets, including also target identity group mentions even when no hate is expressed. Unlike most existing hate speech datasets, which focus on a single target group, our dataset is explicitly designed to capture a diverse range of targets, ensuring a broad and representative sample of hateful (and non-hateful) content. Our work addresses the growing need for updated hate speech datasets, as many existing resources are based on platforms that no longer provide research-friendly data access, such as Twitter (*X*). Crucially, we show that training on existing out-of-domain data leads to poor results on Telegram data, underscoring the necessity of in-domain datasets for effective hate speech detection. We evaluate hate speech classification setups in an extensive series of experiments in both languages, including multilingual, multi-task, and LLM-based approaches. We find that incorporating target information leads to the best performances, enabling multilingual generalization. On the contrary, classification of specific targets shows much room for improvement across setups.

⚠ **Warning**: *this paper contains examples that may be offensive or upsetting.*

Keywords

Telegram, Hate speech, Targets, Polish, Italian

1. Introduction

While a large body of research has focused on hate speech detection in recent years, a significant part of it has been centered on English, especially work that considers different possible targets of hate [1, 2]. Furthermore, while some datasets containing target annotations exist, many of them only focus on one specific kind of hate speech target (e.g., Sanguinetti et al. [3], Bhattacharya et al. [4]).

The most widely used data source in past research for this kind of data has been Twitter (now *X*). However, hate speech detection systems have been found to be subject to performance deterioration when applied to a different domain from the one they were trained on, e.g., a different social network [5, 6] or a different time period [7]. It is therefore important to study different platforms and to develop datasets that can be applied to different use cases. Telegram is an understudied platform compared to Twitter or Facebook, yet it plays a significant role in fringe and extremist communication, especially in light of its anonymity preservation features and reduced content moderation [8].

We present the MuLTA-Telegramdataset, a *Mu*lti-*L*ingual and multi-*Ta*rget dataset developed for the detection of hate speech and its targets on Telegram. It consists of 2,000 messages in Italian and around 2,000 in Polish, annotated for hate speech and its targets, as well as for target identity group mentions.

Crucially, the dataset ensures broad target coverage, as we employed a matrix of keywords to pre-select messages from a large pool of Telegram data and included content representative of 9 minorities target-categories of interest. To ensure that each category is represented across the dataset as a whole and not only within the subset of hateful messages, we annotate the target group mentions, i.e. each message is further assessed on whether its content addresses one or more targets, regardless of whether the message is hateful or not.[1]

Moreover, studying Polish-language content fills a critical gap, given the scarcity of hate speech datasets available and especially given the growing disinformation activity in Central and Eastern Europe [9].

Our aim is that of creating a resource that can be used to train efficient hate speech detection models for textual data, in particular in Italian and Polish, from Telegram, and in the presence of content related to targeted identity groups. After presenting the dataset and its construction, we run a series of experiments under a variety of setups,

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
eleonardelli@fbk.eu (E. Leonardelli); ccasula@fbk.eu (C. Casula); svecelliosalto@fbk.eu (S. Vecellio Salto); jebak@fbk.eu (J. E. Bak); emuratore@fbk.eu (E. Muratore); anna.kolos@nask.pl (A. Kolos); tlouf@fbk.eu (T. Louf); satonelli@fbk.eu (S. Tonelli)

[1]Target mentions and target of hate might not coincide.

including using existing datasets for this task from other social media and LLM annotations, in order to assess the performance of models that are commonly used for this task on our Polish and Italian expert-annotated Telegram data.

The full data and annotations can be obtained at this link: github.com/dhfbk/MuLTa-Telegram.

2. Background

Most existing labeled datasets for abusive language detection are created starting from Twitter (*X*) data, mostly because Twitter data collection APIs were for a long time the easiest to access compared to other platforms [10]. Other less widely used sources for data include Facebook [11, 12] and Instagram [13, 14], while Telegram has been generally overlooked in past work on this topic. Indeed, the only existing resource including hate speech data from Telegram contains automatically-annotated English data from only one Telegram source channel [15], in spite of Telegram having been found to harbor communities that exhibit high levels of toxicity and disinformation across different countries due to its loose data moderation policies [8, 16].

English is the main language represented in existing abusive language datasets [10]. While a number of datasets for detecting abusive language and hate speech in Italian exist, a large number of them consider specific targets or hate-related phenomena, such as racism and xenophobia [17, 3], misogyny [18, 19], religious hate [20], and homotransphobia [21, 22, 23], with some other types of targets often being underrepresented in existing data even for English [24]. Conversely, the available resources for abusive language detection in Polish are rather scarce. The first dataset we could find is described only in a manuscript in Polish from 2017 [25] and it has been publicly available on HuggingFace since 2021.[2] This dataset, however, lacks a detailed description in English. The other available datasets contain posts from Twitter annotated for cyberbullying [26] or offensive comments sourced from a social networking service [27]. We therefore aim at creating a hate speech dataset specifically for Telegram data in Polish and Italian, including expert annotations over 9 categories of identity groups that can be the target of hate.

3. Data Selection and Annotation

In this section, we detail the construction process of our dataset. Public Telegram channels are accessible through a freely available API, originally designed for bot development. While not initially intended for research, this API allows large-scale data collection from public channels. Channels are pages that broadcast self-contained streams of public messages, with posting typically limited to page administrators. Beyond the main chat, channels commonly include additional discussion sections where users can interact with both administrators and one another. We collected data from all these sections.

3.1. Data Collection Strategy

We start from an initial seed set of public Telegram channels known to spread disinformation or hate, curated by a panel of international domain experts in the consortium of the *Hatedemics* European project.[3] As Telegram has a very limited keyword-based search feature, matching only channel titles, we expand these seed channel names using a snowballing approach [28]. This kind of approach consists in first searching for the titles of the seed set channels, and then leveraging Telegram's own user-overlap-based recommendations feature[4] to grow the initial set of channels.

Due to processing constraints, we aim at focusing message retrieval on the most potentially relevant channels for our purpose, identified by the total number of channel recommendations they receive and their distance from seed channels. This distance is defined as the minimum number of recommendation steps required to reach a given channel from a seed. From the top 150 channels in terms of distance from the channels in the seed set and the number of times they were recommended, we retrieve all publicly available messages and associated chat conversations from Jan 1, 2022 to Jan 1, 2023, totaling around 2.5 million messages for Italian and 1.1 million messages for Polish.

3.2. Data Anonymization

With the aim of preserving privacy as much as possible, sensitive information in messages (emails, phone numbers, mentions, etc.) is detected via regular expressions and replaced with placeholders.

Aside from text content, all other information on messages and channels, including channel titles and descriptions, is deleted. This step is carried out to prevent direct identification of the chats in Telegram and to comply with applicable privacy protection regulations.

3.3. Data Pre-Selection for Annotation

Since we aim to detect hateful language in particular across multiple vulnerable social groups, in collaboration with civil society domain experts from NGOs and

[2] https://huggingface.co/datasets/community-datasets/hate_speech_pl

[3] https://hatedemics.eu/

[4] Via GetFullChannelRequest and GetChannelRecommendation in Telethon: https://github.com/LonamiWebs/Telethon

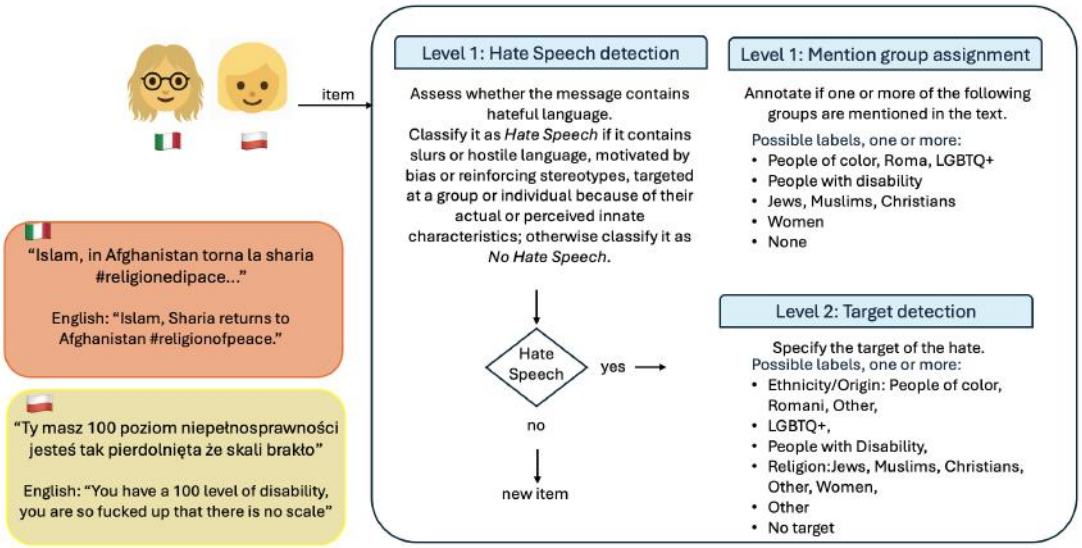

Figure 1: Annotation Scheme and examples taken from the dataset.

research institutions, we have defined a set of common targets of hate in the countries and contexts we take under consideration, including People with Disabilities; LGBTQ+ Individuals; Religion: Jews, Muslims, Christians; Ethnicity/Origin: People of Color, Romani people, Other (including Migrants); Women. These target identity groups have been partially adapted from the ones used in the Measuring Hate Speech corpus [1, 2], which uses US-centric identity categories, adjusting them to our European context.

We then developed a keyword matrix consisting of 145 group-specific terms.[5] These keywords have been selected based on prior domain expertise and preliminary corpus exploration.

Aiming at obtaining a high representation of content related to the target identity groups we identified, we then carried out a pre-selection step. From the entire Telegram data collection, we pre-selected for manual annotation about 1,500 posts (75% of the entire dataset) containing at least two distinct keywords (from our matrix) associated with the same target group. This is done using a string-matching filter. We then construct the remaining 25% of the dataset by randomly selecting posts to manually annotate, in order to create a more representative overall sample of random messages on Telegram, which of course might not contain target-related words.

[5]The keyword matrix is available on github: https://github.com/dhfbk/MuLTa-Telegram.

3.4. Data Annotation

We employ expert Polish and Italian annotators, two Italian native speakers (one male, age 26, and one female, age 41) and two Polish native speakers (one female, age 22, and one female, age 37). Annotators were asked to indicate whether a message contained hate speech. If hate speech was present, annotators were required to specify the target of the hate speech from our predefined list of categories. To gain a deeper understanding of the dataset's content and to ensure that the dataset covered a broad range of target identity categories not only in the hateful part of the dataset, annotators were also asked to label the target mentions of each message among a set of predefined categories.[6] An overview of the annotation scheme that was used for annotating both the Italian and the Polish sections of the dataset is provided in Figure 1, while the full annotation guidelines are reported in Appendix 8.1. This process resulted in two comprehensive databases containing messages annotated for both hateful and non-hateful content, targeting various identity groups. A numerical breakdown of their content is provided in Table 1, 2 and in Figure 2.
The databases mainly contain non-hateful messages, with the Italian one featuring almost as many hate messages as the Polish one. This may be due to different use of Telegram or to a greater number of controversial, yet not explicitly hateful, messages in the Polish database, which includes many discussions related to the Russia-Ukraine

[6]Target mentions assignation and target of hate might not coincide.

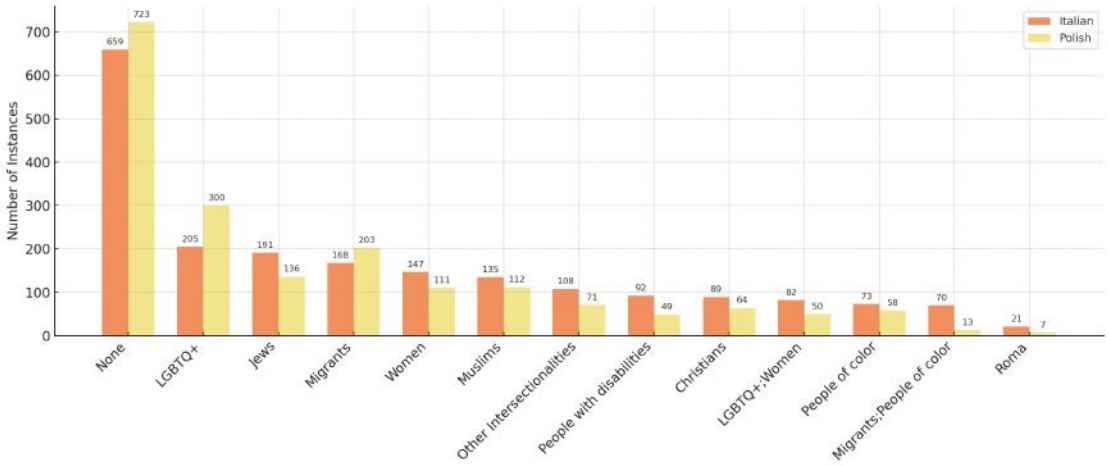

Figure 2: Mentioned group distribution in the datasets.

war. When analyzing the targets of hate speech, most messages are directed at ethnic groups, with a prevalence of attacks against people of color in Italian and against Ukrainian refugees in Polish, followed by those targeting LGBTQ+ identities. While a significant portion of hateful messages targets either groups not represented in the selected taxonomy (*Other*) or expresses hate without a specific target (*No Target*), there is little representation of hate toward the remaining identity categories.

Table 1
Statistics of the manually annotated datasets.

		N. Messages
Italian	Total Messages	2,002
	Hate Speech	411 (20.5%)
Polish	Total Messages	1,934
	Hate Speech	249 (12.9%)

Table 2
Statistics of the targets of hate speech.

Target	Italian	Polish
LGBTQ+	79	49
Ethnicity/Origin: Other	39	73
Ethnicity/Origin: POC	99	8
Religion: Jewish	13	14
Women	26	6
Ethnicity/Origin: Romani	8	1
Religion: Muslims	3	4
People with Disability	4	0
Religion: Christians	1	2
Other	115	55
No Target	24	37
Total Hateful messages	411	249

3.5. Inter-Annotator Agreement

Inter-annotator agreement was calculated for each language on a sub-sample of 200 posts using Krippendorff's alpha, annotated each by two expert annotators who are native speakers of Italian or Polish. The Polish portion of the dataset showed an IAA of 0.41, while the Italian one 0.68. These numbers, while low, are in line with previous work on similar topics, especially considering that our annotators had no chance to discuss and revise their annotations together, as they worked asynchronously. For instance, Basile et al. [29] showed an inter-rater agreement for aggressiveness in Spanish of 0.47.

4. Classification Experiments

As a way to benchmark our newly-created dataset, and to explore different strategies for classification of hate speech in Italian and Polish on Telegram data, we devise a series of experiments using different experimental setups. These experiments include fine-tuning BERT-base classifiers (Sec. 4.1), multi-task models (Sec. 4.2), and LLM prompting (Sec. 4.3). To evaluate approaches across different experiments in a comparable way, 35% of the manually annotated dataset was withheld and used as test set for each language. The remaining 1,300 manually annotated items (65%) were used to fine-tune models where necessary (i.e., Experiments 2 and 4). Each experiment was replicated with a consistent setup across both languages.

4.1. Supervised Hate Speech Detection via BERT Fine-Tuning

In this set of experiments we fine-tune existing monolingual (Exp. 1,2,3) and a multilingual (Exp.4) BERT-based language models [30].
Regarding monolingual models, for Polish we conducted a series of experiments using three distinct BERT-based models for the Polish language: we used a general-purpose Polish BERT-model (*BERT-base-pl*)[7] and two models trained for identifying specific types of offensiveness, namely cyberbullying (*BERT-cb-pl*)[8] and hate speech (*BERT-hs-pl*).[9]

For Italian, we fine-tuned a general-purpose Italian BERT-based model (*BERT-base-it*),[10] a BERT-based model pre-trained on Italian data from Twitter (*AlBERTo*) [31],[11] and a binary hate speech classification model for Italian social media text (*Hate-ita*) [32].[12]

For fine-tuning the models we employed the MaChAmp library [33], an open-source tool designed to simplify flexible tasks configuration, multitask and multilingual fine-tuning of transformer-based language models. All the evaluated models were fine-tuned for 5 epochs using a single GPU, applying the default hyperparameters provided by MaChAmp (see Appendix 8.2). To address class imbalance, we assign equal weight to each class during training, ensuring that minority classes are not underrepresented.

Experiment 1: Training on Existing Datasets Our first experiment aims to evaluate the performance of models fine-tuned on other publicly available datasets on our manually-annotated Telegram test data. They serve as a baseline.

For Italian, we use 2,000 examples from 4 existing datasets that represent some of the targets we consider in our work: the AMI dataset [34], focused on misogyny; the Haspeede dataset [35], focused on hateful content against Muslims, immigrants and Roma people; the HODI dataset [23], a dataset for detection of homotransphobia in Italian; and the Religous Hate dataset [36], an Italian dataset that includes Anti-Judaism, Anti-Christianity and anti-Islam social media posts.[13]

For Polish, we could find 3 datasets total related to online abusive content. We decided to discard the oldest one [25] due to lack of available information on its construction (data collection, annotation, content) and because after a preliminary manual inspection our annotators found the data to be noisy (e.g., HTML code was found in the middle of the texts).

This left us with two datasets for hate speech, which we use in combination in our experiments: the Cyberbullying dataset [26] and the BAN-PL dataset [27]. These datasets differ significantly in both their definitions of hate and their annotation procedures. For instance, the Cyberbullying dataset contains generally milder or less severe phenomena in its annotations, as it is focused on the somewhat broader phenomenon of cyberbullying compared to hate speech. In contrast, BAN-PL considers a message as Not Hateful if it remained online for more than two days without being removed by a platform moderator. Only a small subset of the removed comments was then manually annotated as Hateful.

Given these differences, we opted to use only the manually annotated hateful samples from BAN-PL, which are more aligned with our definition of hate speech. For the neutral (non-hateful) class, we combined equal portions of BAN-PL and Cyberbullying data, ensuring a balanced yet representative dataset composition.

Experiment 2: Fine-tuning on Manually Annotated Data This is the main experiment in which we evaluate the potential usefulness of our dataset for training hate speech detection models. We fine-tune the models on 1,300 manually annotated items from our dataset for each language. The task setup is single-task, focusing exclusively on the hate speech task. Since the annotated data is in-domain, we expect this setup to yield better performance on our Telegram test data compared to Experiment 1, which used out-of-domain data (i.e., data from different platforms).

Experiment 3: Fine-tuning on LLM-Annotated Data (LLaMA) To investigate whether LLMs can serve as a viable alternative to manual annotation in hate speech detection tasks on Telegram, we devise an experiment in which we use LLaMA 3.1 70B Instruct as an automated annotator. We ask the model to annotate the same train split of our dataset as in Experiment 2, by prompting the model with a summary of our hate speech annotation guidelines. For both languages, we then fine-tune the same BERT-based models as in Experiment 2, but this time on the LLM-annotated data. We then evaluate the trained models on the test sets.

Experiment 4: Multilingual BERT A multilingual approach can leverage shared representations across languages. In this context, a model is required to generalize patterns that may be strongly language- and context-dependent, a non-trivial task. Nonetheless, this strategy offers several advantages: it can boost performance in

[7] dkleczek/bert-base-polish-uncased-v1
[8] ptaszynski/bert-base-polish-cyberbullying
[9] dkleczek/Polish-Hate-Speech-Detection-Herbert-Large
[10] dbmdz/bert-base-italian-cased
[11] m-polignano-uniba/bert_uncased_L-12_H-768_A-12_italian_alb3rt0
[12] MilaNLProc/hate-ita
[13] Given that this dataset contains several targets in addition to religion-focused ones, we filtered it to retain only religious targets.

low-resource settings through cross-lingual transfer, and it can improve robustness by exposing the model to more diverse inputs during training.

To test the viability of this approach, we merge the two manually annotated train splits of the Polish and Italian datasets to fine-tune a multilingual BERT base model.[14] The performance of the model for classification of hate speech is then evaluated on the Italian and Polish test sets separately.

4.2. Multi-task Setup for Hate Speech and Target Detection

Experiment 5 Given the hierarchical relationship between hate speech detection and target identification, we adopt a multi-task learning approach to jointly model these tasks, under the assumption that each task can help generalization on the other. In this multi-task learning paradigm, schematically illustrated in Table 3, the model can jointly optimize for different tasks, allowing all tasks to benefit from shared signals captured through a common representation, which is jointly fine-tuned during training. This approach is motivated by prior work showing that training models on related tasks simultaneously can lead to better performance than training them in isolation [37]. This setup should allow to improve generalization and stability of the hate speech task, but also to automatically predict the targets of hate speech, a task that as a single task would be extremely difficult to address with the currently available data, given the scarcity of targets (see Table 3.4).

In this setting, hate speech detection serves as the primary task, since the presence of a target group in a message depends on the detection of hate speech in the first place, while target identification is treated as a secondary task. Specifically, we used our pre-trained models as the shared encoder for both tasks, while a separate decoder is utilized by each task. We incorporate different loss weighting to the two tasks, in order to represent the hierarchy of primary and auxiliary.[15]

4.3. Prompt-Based Hate Speech Detection via LLMs

Experiments 6 and 7: Llama We then aim at evaluating the performance of LLMs on our Telegram annotated data in Italian and Polish. For this, we use LLaMA [38], since it possesses some multilingual capabilities, especially in Italian. In particular, we prompt LLaMA 3.1 70B Instruct (Exp. 6) with our annotation guidelines and ask it to label each test example as hateful or not. We then also evaluate LLaMA Guard (Exp. 7), using no prompt as it is a model explicitly made to detect inappropriate or toxic content.[16]

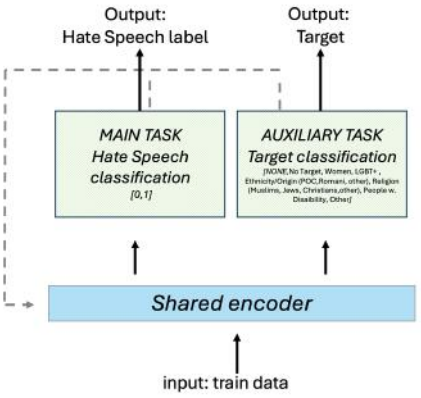

Figure 3: The design of the multitask setup used for experiment 5.

While this kind of experimental setup is useful for comparison purposes, it should be noted that it is highly inefficient, and unlikely to be feasible and scalable when large amounts of data need to be processed at once, as its computational speed and efficiency is much lower than that of a BERT-based model fine-tuned on task-specific data. Such models are particularly well-suited for social science research, where cost-effective processing of millions of messages is often required to study trends in online hate and its societal impact. Given that our goal is the development of hate speech classification models that can be employed in real-life scenarios, we consider LLM-based classification out of this scope.

5. Experimental Results and Discussion

In this section, we present the results obtained in our experiments. A summary of the results across all experimental setups is shown in Tables 3 and 4. For the experiments using multiple models (Exp. 1, 2, 3, and 5), we report average macro-F_1 scores, while the detailed results are in Appendix 8.3. As a first general observation, Polish and Italian show consistent results patterns across experiments, which allows us to derive meaningful observations across both languages.

[14] google-bert/bert-base-multilingual-cased

[15] The multi-task learning loss is computed as $L = \sum_t \lambda_t L_t$, where L_t is the loss for task t and λ_t the corresponding weighting parameter, and we provide a different loss weight for the auxiliary tasks. For the main task, we empirically set $\lambda_t = 0.7$, and $\lambda_t = 0.3$ for the auxiliary task.

[16] https://huggingface.co/meta-llama/Llama-Guard-3-8B

Table 3
Summary and F1 scores for Italian and Polish across experimental setups. For Experiments 1-3 and 5, the average F1 across the three used models is shown. Full results in Appendix.

Exp.	model(s)	trained on	annotation	setup	Italian F1	Polish F1
Exp1	BERT-based models	out-domain	mixed	single task	0.672 ± 0.015	0.50 ± 0.018
Exp2	BERT-based models	in-domain	manually	single task	0.717 ± 0.072	$\mathbf{0.846 \pm 0.007}$
Exp3	BERT-based models	in-domain	Llama	single task	0.705 ± 0.017	0.658 ± 0.015
Exp4	multilingual BERT	in-domain	manually	single task	0.589	0.564
Exp5	BERT-based models	in-domain	manually	multitask	$\mathbf{0.732 \pm 0.025}$	0.801 ± 0.016
Exp6	Llama	-	-	prompted	**0.732**	0.678
Exp7	LLama-Guard	-	-	no prompt	0.712	0.58

5.1. Hate Speech Detection

The results of the binary classification of hate speech are reported in Table 3. In-domain training (Exp. 2) consistently outperforms the models trained on out-of-domain data (Exp. 1) across both languages, underscoring the necessity of domain-specific data. Notably, out-of-domain training results in the worse classification performance for Polish and the second worse for Italian.

Conversely, the training of multilingual BERT (Exp. 4) resulted in very low performance overall, suggesting that models trained across multiple languages can struggle to generalize effectively for this task. Regarding specific model performances, for both languages, fine-tuning a model already fine-tuned for hate speech (`Hate-ita` and `BERT-hs-pl`, for Italian and Polish respectively) leads to the best results within models across all experiments.[17]

The Llama-based experiments, including Exp. 3, in which Llama was used to annotated data for training a BERT-based classifier, and Exps. 6 and 7, in which Llama (70B Instruct and Llama Guard) predicted test set labels through prompting, yielded intermediate performance.

While generally better than out-of-domain approaches, they consistently fell short of models trained on expert human annotations. Llama-based predictions performed consistently worse in the case of Polish, possibly due to the model lacking official support for the Polish language.

The multi-task setup (Exp. 5), on the other hand, improved hate speech detection performance, achieving the highest macro-F_1 scores for both languages.

5.2. Target Identification

Regarding the parallel task of target of hate identification, while overall performances appear high in both languages (Accuracy: Polish 87%, Italian 82%), this result is driven primarily by the model's strong performance on the majority class, i.e., samples in the *non-hate* class, therefore without target, which heavily skews the results. Macro-averaged F_1 scores on each target are very low, as shown in Table 4, indicating very poor performance on minority classes prediction (hateful and targeted examples). Notably, for Italian the most frequent target class *Ethnicity/Origin: Person of Color* is consistently recognized (with an F1-score of almost 0.70), and performance on the moderately frequent class *LGBTQ+* depends on the model (F_1 scores range from 0.00 to 0.41), while the other target groups are entirely or almost entirely disregarded. For Polish, the target *LGBTQ+* is classified more accurately than the others (F1 0.29 up to 0.69).

Table 4
Average F1-scores (across the three evaluated Bert-based models) per target category in Italian and Polish.

Target	Italian	Polish
LGBTQ+	0.24±0.21	0.52±0.21
Ethnicity/Origin: Other	0.00	0.28±0.06
Ethnicity/Origin: PoC	0.66± 0.08	0.00
Religion: Jewish	0.00	0.00
Women	0.00	0.00
Ethnicity/Origin: Romani	0.00	0.10±0.17
Religion: Muslim	0.00	0.00
People w. Disabilities	0.00	0.00
Religion: Christians	0.00	0.00
Other	0.13±0.15	0.28±0.07
No Target	0.00	0.44±0.08
NONE (no HS)	0.91±0.03	0.95±0.00

5.3. Additional Multilingual Experiments

Given the very low performance of the multilingual model (Italian: 0.589, Polish: 0.564 F1), we sought to investigate potential causes for this. Although different languages might express hate differently, and context can vary, one possible factor that could explain the low performance of multilingual models is annotation inconsistencies between the Italian and Polish datasets, especially given the difficulty and subjectivity of the type of annotation.

To investigate this, we repeated Experiment 4 by fine-tuning multilingual BERT, this time using the data from Experiment 3, which was annotated via LLM. These LLM-

[17] For more detailed results see Appendix 8.3.

generated annotations should in principle be more homogeneous across languages, assuming the system is using the same criteria given the same prompt instructions. In this setup, performance improved notably (Italian: 0.674, Polish: 0.657 F1), supporting our hypothesis.

Nonetheless, we were interested in performance of our the best performing scenario, i.e. on high-quality, manually annotated data and multitask setup. We re-ran the experiment using multitask learning (i.e., jointly predicting hate speech and its target) on the human-labeled datasets. This yielded the best results for both languages (Italian: 0.706, Polish: 0.726 F1).

These findings suggest that inconsistencies among annotators across languages can hamper results of multilingual models, but learning on richer data can help, since an auxiliary task can help generalization by providing more training signal and regularizing the model.

6. Manual Qualitative Analysis

To understand the differences between the Italian and Polish data, we conducted a manual qualitative analysis. First, we noticed a disparity in the distribution of hateful messages targeting *Ethnicity/Origin*. While the Italian dataset shows a predominance of messages directed at people of color (99 instances, compared to 9 in the Polish dataset), the subcategory *Other (Migrants)* appears less frequently in Italian (39 instances) than in Polish (82). These patterns likely reflect the socio-political context at the time of data collection, with immigration by people of color being a prominent issue in Italy and the presence of Ukrainian refugees being central in Poland. This underscores the importance of collecting context-sensitive data, particularly at the socio-cultural level, as each context can exhibit different patterns and phenomena.

We also investigated the discrepancies between automatic prediction and human annotation. We identified 29 Italian messages and 30 Polish ones which the annotators deemed hateful and the models classified otherwise. For the opposite case, there were 70 messages in Italian and only 3 messages in Polish.

In the first case, models seem unable to detect hateful content when not presented in a standard explicitly offensive form. Performance tends to be low when examples include hashtags ("***Islam**, in Afghanistan torna la sharia [...] **#religionedipace...***" ["*Islam, in Afghanistan sharia is back [...] #religionofpeace..*"]); dehumanization being implied (*"i roma [...]* ***non sono*** *veri* ***esseri umani****, punto"* ["*the Roma [...] are not real human beings, period*"], *"**I kulka** we własny łeb*" ["*And a bullet to your own head*"]); slurs in non-standard language varieties (*"**Na Zengara** in pratica" ["*Basically about a gypsy*"]*); and occasionally established slurs (e.g., Italian n-word). Models appear less proficient than humans in detecting implied hate speech, especially in the absence of profanity.

In the second case, models overestimated hatred in messages expressing controversial opinions (*"non c'è nessun isolamento perché **non esistono i virus**" ["There's no isolation because viruses don't exist"], "**Lepiej dla Ruskich**, kto lubi ten shit?"* ["*Better for the Russians, who likes that shit?*"]) or sensitive topics (*"**Una pacca sul sedere** non autorizzata è una molestia sessuale"* ["*An unwarranted slap on the butt is sexual harassment*"]). Additionally, models struggled with relatively mild insults containing no targets in the given context (*"Nikt nie pomoże.. **Bandyci** bezkarni.."* ["*No one will help.. Bandits unpunished..*"]), idiomatic use of expressions related to disabilities, which are lexicalized in spoken Italian, albeit unkind (*"purtroppo non c'è peggior **sordo** di chi non vuol sentire e peggior **cieco** di chi non vuol vedere"* ["*Unfortunately, there's no one more deaf than those who don't want to hear, and no one more blind than those who don't want to see*"]), or critiquing hateful messages (*"tipico cristiano ipocrita...va in chiesa però vorrebbe **sterminare** chi crede nel Islam"* ["*Typical hypocritical Christian...goes to church but would like to exterminate those who believe in Islam*"]). Finally, some cases appear to be simply annotation errors (*"@<user> finalmente Instagram mi dà le **pubblicità** giuste"* ["*@<user> finally Instagram shows me the right ads*"]).

7. Conclusions

In this paper, we introduced MuLTa-Telegram, a novel multilingual dataset for hate speech and target detection, containing data from Telegram in both Italian and Polish.

The dataset includes anotations across 9 hate speech target categories, in contrast with the majority of available datasets, which are often limited to single targets. Moreover, we ensured the presence of target-related content also in the non-hateful part of the dataset, with about 75% of the messages containing target-relevant content (see Figure 2). Furthermore, while the vast majority of hate speech research has been conducted on English, we focused on underrepresented languages.

We conducted an extensive set of experiments, showing that the fine-tuning of BERT-based models on out-of-domain hate speech classification data leads to poor performance on Telegram data, while training on in-domain resources consistently outperforms it. This draws attention to the limitations of relying on datasets from platforms like Twitter, which are no longer reliably accessible for academic research, reinforcing the need for updated and diversified resources like MuLTa-Telegram. However, results on the detection of individual targets remained poor, particularly for more scarcely represented groups. This underscores the persistent difficulty of detecting hate directed at less-represented communities.

Furthermore, in a multilingual setup, we showed how the addition of a parallel task predicting targets greatly improves performances for hate speech classification, enabling the model to generalize across languages. We included both LLaMA and LLaMA Guard in our evaluation to explore how general-purpose and safety-focused systems perform on our task. LLaMA Guard, despite its safety orientation, performs poorly in this out-of-domain context, while LLaMA shows strong performance on Italian, but its accuracy drops on Polish data, likely due to limited language coverage during pretraining. These results emphasize the need for both domain- and language-specific adaptation.

While we fine-tuned transformer-based models directly on classification tasks using Telegram data, future work could explore domain-adaptive pretraining via Masked Language Modeling on unlabeled Telegram messages. This step could improve the encoder's alignment with the linguistic characteristics of the platform, potentially enhancing classification performance.

We hope this dataset will help foster research into hate speech detection for underrepresented languages and platforms. Future work will explore expanding the dataset to more languages and domains, as well as improving the detection of fine-grained targets of hate.

Acknowledgments

The work of C. Casula, A. Kołos, E. Leonardelli, E. Muratore and S. Tonelli has been supported by the European Union's CERV fund under grant agreement No. 101143249 (HATEDEMICS). This research was also partially supported by the European Union under the Horizon Europe project AI-CODE, GA No. 101135437.

References

[1] C. J. Kennedy, G. Bacon, A. Sahn, C. von Vacano, Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application, 2020. URL: http://arxiv.org/abs/2009.10277. doi:10.48550/arXiv.2009.10277, arXiv:2009.10277 [cs].

[2] P. Sachdeva, R. Barreto, G. Bacon, A. Sahn, C. von Vacano, C. Kennedy, The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism, in: G. Abercrombie, V. Basile, S. Tonelli, V. Rieser, A. Uma (Eds.), Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, European Language Resources Association, Marseille, France, 2022, pp. 83–94. URL: https://aclanthology.org/2022.nlperspectives-1.11.

[3] M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, M. Stranisci, An {I}talian Twitter Corpus of Hate Speech against Immigrants, in: Proceedings of the 11th Language Resources and Evaluation Conference, European Language Resources Association, Miyazaki, Japan, 2018, pp. 2798–2805.

[4] S. Bhattacharya, S. Singh, R. Kumar, A. Bansal, A. Bhagat, Y. Dawer, B. Lahiri, A. K. Ojha, Developing a multilingual annotated corpus of misogyny and aggression, in: R. Kumar, A. K. Ojha, B. Lahiri, M. Zampieri, S. Malmasi, V. Murdock, D. Kadar (Eds.), Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 158–168. URL: https://aclanthology.org/2020.trac-1.25.

[5] J. Salminen, M. Hopf, S. A. Chowdhury, S.-g. Jung, H. Almerekhi, B. J. Jansen, Developing an online hate classifier for multiple social media platforms, Human-centric Computing and Information Sciences 10 (2020) 1. URL: https://doi.org/10.1186/s13673-019-0205-6. doi:10.1186/s13673-019-0205-6.

[6] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, S. Villata, Cross-platform evaluation for italian hate speech detection, in: CLiC-it 2019-6th Annual Conference of the Italian Association for Computational Linguistics, 2019.

[7] K. Florio, V. Basile, M. Polignano, P. Basile, V. Patti, Time of your hate: The challenge of time in hate speech detection on social media, Applied Sciences 10 (2020). URL: https://www.mdpi.com/2076-3417/10/12/4180. doi:10.3390/app10124180.

[8] R. Rogers, Deplatforming: Following extreme internet celebrities to telegram and alternative social media, European Journal of Communication 35 (2020) 213–229. doi:10.1177/0267323120922066.

[9] M. Wenzel, K. Stasiuk-Krajewska, V. Macková, K. Turková, The penetration of russian disinformation related to the war in ukraine: Evidence from poland, the czech republic and slovakia, International Political Science Review 45 (2024) 192–208. doi:10.1177/01925121231205259.

[10] B. Vidgen, L. Derczynski, Directions in abusive language training data, a systematic review: Garbage in, garbage out, PLOS ONE 15 (2020) e0243300. doi:10.1371/journal.pone.0243300.

[11] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, M. Tesconi, Hate me, hate me not: Hate speech detection on Facebook, in: Italian Conference on Cybersecurity, 2017.

[12] R. Kumar, A. N. Reganti, A. Bhatia, T. Maheshwari, Aggression-annotated corpus of Hindi-English code-mixed data, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceed-

ings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1226.

[13] F. A. Vargas, I. Carvalho, F. R. de Góes, F. Benevenuto, T. A. S. Pardo, Building an Expert Annotated Corpus of Brazilian Instagram Comments for Hate Speech and Offensive Language Detection, arXiv:2103.14972 [cs] (2021). URL: http://arxiv.org/abs/2103.14972, arXiv: 2103.14972.

[14] V. Parvaresh, Covertly communicated hate speech: A corpus-assisted pragmatic study, Journal of Pragmatics 205 (2023) 63–77. URL: https://www.sciencedirect.com/science/article/pii/S037821662200296X. doi:https://doi.org/10.1016/j.pragma.2022.12.009.

[15] V. Solopova, T. Scheffler, M. Popa-Wyatt, A telegram corpus for hate speech, offensive language, and online harm, Journal of Open Humanities Data (2021). doi:10.5334/johd.32.

[16] A. Urman, S. Katz, What they do in the shadows: examining the far-right networks on telegram, Information, Communication & Society 25 (2022) 904–923. URL: https://doi.org/10.1080/1369118X.2020.1803946. doi:10.1080/1369118X.2020.1803946.

[17] C. Bosco, V. Patti, S. Frenda, A. T. Cignarella, M. Paciello, F. D'Errico, Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP, Information Processing and Management 60 (2023) 103118. URL: https://www.sciencedirect.com/science/article/pii/S0306457322002199. doi:https://doi.org/10.1016/j.ipm.2022.103118.

[18] P. Zeinert, N. Inie, L. Derczynski, Annotating Online Misogyny, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3181–3197. doi:10.18653/v1/2021.acl-long.247.

[19] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, H. Margetts, An Expert Annotated Dataset for the Detection of Online Misogyny, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1336–1350.

[20] A. Ramponi, B. Testa, S. Tonelli, E. Jezek, Addressing religious hate online: from taxonomy creation to automated detection, PeerJ Computer Science 8 (2022) e1128.

[21] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for identification of homophobia and transophobia in multilingual youtube comments, 2021. arXiv:2109.00227.

[22] D. Locatelli, G. Damo, D. Nozza, A cross-lingual study of homotransphobia on twitter, in: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), 2023, pp. 16–24.

[23] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, Hodi at evalita 2023: Overview of the first shared task on homotransphobia detection in italian 113 (2024) 26.

[24] C. Casula, S. Tonelli, On the Impact of Hate Speech Synthetic Data on Model Fairness, in: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), 2025.

[25] M. Troszyński, A. Wawer, Czy komputer rozpozna hejtera? wykorzystanie uczenia maszynowego (ml) w jakościowej analizie danych, Przegląd Socjologii Jakościowej 13 (2017) 62–80.

[26] M. Ptaszynski, A. Pieciukiewicz, P. Dybala, P. Skrzek, K. Soliwoda, M. Fortuna, G. Leliwa, M. Wroczynski, Expert-Annotated Dataset to Study Cyberbullying in Polish Language, Data 9, 1 (2024). URL: https://doi.org/10.3390/data9010001. doi:10.3390/data9010001.

[27] A. Kolos, I. Okulska, K. Głąbińska, A. Karlińska, E. Wiśnios, P. Ellerik, A. Prałat, Ban-pl: A polish dataset of banned harmful and offensive content from wykop. pl web service, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 2107–2118.

[28] J. Baumgartner, S. Zannettou, M. Squire, J. Blackburn, The pushshift telegram dataset, in: Proceedings of the international AAAI conference on web and social media, volume 14, 2020, pp. 840–847.

[29] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. doi:10.18653/v1/S19-2007.

[30] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[31] M. Polignano, P. Basile, M. De Gemmis, G. Semeraro, V. Basile, et al., Alberto: Italian bert language understanding model for nlp challenging tasks based on

tweets, in: CEUR workshop proceedings, volume 2481, CEUR, 2019, pp. 1–6.
[32] D. Nozza, F. Bianchi, G. Attanasio, Hate-ita: Hate speech detection in italian social media text, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), 2022, pp. 252–260.
[33] R. Van Der Goot, A. Üstün, A. Ramponi, I. Sharaf, B. Plank, Massive choice, ample tasks (machamp): A toolkit for multi-task learning in nlp, arXiv preprint arXiv:2005.14672 (2020).
[34] E. Fersini, D. Nozza, P. Rosso, et al., Ami@ evalita2020: Automatic misogyny identification, in: Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), (seleziona...), 2020.
[35] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2@ evalita2020: Overview of the evalita 2020 hate speech detection task, Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2020).
[36] A. Ramponi, B. Testa, S. Tonelli, E. Jezek, Addressing religious hate online: from taxonomy creation to automated detection, PeerJ Computer Science 8 (2022) e1128.
[37] R. Caruana, Multitask learning, Machine learning 28 (1997) 41–75.
[38] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
[39] R. van der Goot, A. Üstün, A. Ramponi, I. Sharaf, B. Plank, Massive choice, ample tasks (machamp): A toolkit for multi-task learning in nlp, 2021. URL: https://arxiv.org/abs/2005.14672. arXiv:2005.14672.

8. Appendices

8.1. Annotation Guidelines

In this section we report the annotation guidelines.

Hate Speech Detection

Assess whether the message contains hateful language. Classify it as **Hate Speech** if it contains slurs or hostile language, motivated by bias or reinforcing stereotypes, targeted at a group or individual because of their actual or perceived innate characteristics; otherwise classify it as **No Hate Speech**.

- Reported speech is **not** hate speech.
- Text can be hateful even if the target is implicit, as long as it's implied by the context.
- It is **not** hateful if the target is an organization and not its members.
- Profanities alone do not imply hatefulness, unless the tone is aggressive or the message is clearly directed toward someone (e.g., *"Aspetta che li **minacciano** per bene e poi vedi se accettano..."*).
- False or debatable statements do not imply hatefulness, but messages that erase identities (e.g., *"esistono **solo** due sessi"*) **are** hate speech.
- References to individuals or citizens (excluding military groups) as *nazis* in the context of the Russia-Ukraine war are to be considered hate speech.
- In Polish:
  - If *"Banderowiec"* refers to supporters of Stepan Bandera (OUN), it is not hate speech.
  - If *"Banderowiec"* is used to refer to the entire Ukrainian nation or other social groups in a hateful or offensive way, it is hate speech.

Target Detection

When text contains hate speech, specify its target. Possible categories include:

- *Ethnicity/Origin: People of Color, Romani, or Other (Migrants)*
- *LGBTQ+*
- *People with Disability*
- *Religion: Jewish, Christians, Muslims, Other*
- *Women*
- *Other*
- *No Target*

Choose the most appropriate category. Select *other* for any specific target not included in any other category. Select *No Target* for occurrences of hate speech not directed at any specific group.

- In cases where multiple labels apply, prioritize the identity that is most harmed.
- The target must be explicitly addressed, not implied (e.g., by referring to stereotypical associations):
  - Talking about Arabic/Muslim countries or Islam does **not** imply the *Muslim* target.
  - Talking about Africa or African migration does **not** imply the *People of Color* target.
  - Mentions of disability imply the *People with Disability* target.

- If references to disability or any identity group are used idiomatically or as insults, label them as idiomatic.
- If the word *woman* is mentioned as one of the sexes or if the subject is a specific woman, select the target *Women*.

Mention of Target Group Detection

Annotate if one or more of the following groups are addressed in the text. Assign the corresponding label(s). Multiple groups may be annotated for a single message. Possible target groups include:

- *Ethnicity/Origin: People of color, Romani, Other (Migrants)*
- *LGBTQ+*
- *People with Disability*
- *Religion: Jews, Muslims, Christians, Other*
- *Women*
- *None*

If none of these target groups are addressed, assign the label *None*. A group should be annotated if it is explicitly mentioned or implicitly clear from the context. Annotate a group even if it is not the main focus of the message.

8.2. Hyperparameters

In this section we described the parameters used for BERT-based experiments.

Table 5
Default MaChAmp hyperparameter settings [39] used for all our experiments.

Hyperparameter	Value
Optimizer	AdamW
β_1, β_2	0.9, 0.99
Dropout	0.3
Epochs	10
Batch size	32
Learning rate (LR)	0.0001
LR scheduler	Slanted triangular
Decay factor	0.38
Cut fraction	0.2

8.3. Experimental Results

In this section, in Tables 6 and 7 for Italian and Polish respectively, we report the full detailed results for Experiments 1,2,3 and 5.

Table 6
F1 Scores Across Experiments for Hate Speech Detection Models for Italian.

Experiments	Model	Italian Macro		Non-Hate		Hate	
		F1	Avg	F1	avg	F1	avg
Exp1 - out of domain data	AlBERTo	0.658	0.672 ±0.015	0.804	0.813±0.028	0.512	0.53 ±0.018
	BERT-base-it	0.688		0.844		0.531	
	Hate-ita	0.669		0.79		0.548	
Exp2 - manually annotated data	AlBERTo	0.758	0.717±0.072	0.911	0.909±0.004	0.605	0.525±0.14
	BERT-base-it	0.634		0.905		0.363	
	Hate-ita	**0.759**		0.912		0.607	
Exp3 - Llama as annotator	AlBERTo	0.686	0.705±0.017	0.816	0.844±0.03	0.556	**0.566±0.014**
	BERT-base-it	0.718		0.875		0.561	
	Hate-ita	0.711		0.84		0.582	
Exp 4 - multilingual	BERT-multilingual	0.589	-	0.896	-	0.282	
Exp5 - multitask setup	AlBERTo	0.703	**0.732±0.025**	0.907	**0.912±0.005**	0.5	0.551±0.045
	BERT-base-it	0.743		0.915		0.571	
	Hate-ita	0.749		0.915		0.582	
Exp 6 - Llama	LlaMA 3.1 70B Ins.	0.732	-	0.852	-	0.613	-
Exp 7 - Llama Guard	Llama-Guard-3-8B	0.712	-	0.862	-	0.561	-

Table 7
F1 Scores Across Experiments for Hate Speech Detection Models for Polish.

Experiments	Model	Polish Macro		Non-Hate		Hate	
		F1	Avg	F1	avg	F1	avg
Exp1 - out of domain data	BERT-base-pl	0.472	0.50±0.018	0.606	0.637±0.02	0.337	0.362±0.013
	BERT-hs-pl	0.561		0.714		0.408	
	BERT-cb-pl	0.466		0.592		0.34	
Exp2 - manually annotated data	BERT-base-pl	0.833	**0.846±0.01**	0.955	0.96±0.002	0.71	0.73±0.013
	BERT-hs-pl	0.835		0.96		0.779	
	BERT-cb-pl	0.871		0.964		0.779	
Exp3 - Llama as annotator	BERT-base-pl	0.606	0.658±0.015	0.79	0.84±0.015	0.422	0.476±0.016
	BERT-hs-pl	0.698		0.876		0.52	
	BERT-cb-pl	0.671		0.855		0.488	
Exp 4 - multilingual	BERT-multilingual	0.564	-	0.926	-	0.202	-
Exp5 - multitask setup	BERT-base-pl	0.797	0.80±0.015	0.951	0.95±0.003	0.642	0.65±0.03
	BERT-hs-pl	0.755		0.941		0.568	
	BERT-cb-pl	0.85		0.959		0.742	
Exp 6 - Llama	LlaMA 3.1 70B Ins.	0.678	-	0.843	-	0.512	-
Exp 7 - Llama Guard	Llama-Guard-3-8B	0.58	-	0.814	-	0.347	-