# Elementi per un progetto sul tema della creazione di dataset su Hate Speech misogino in lingue non ancora disponibili

The ultimate goal of the project is to serve the cause of gender equality by increasing the number and quality of sentiment analysis tools able to monitor misogynous speech.

This goal is pursued through the analysis of the current initiatives aiming at the creation of datasets on misogynous hate speech for underrepresented languages, and proposing an innovative partnership aiming at boosting the process.

**Theory of change**

* If the corpus of datasets for misogynous hate speech detection still overlooks a large number of languages
* If current private and public initiatives tend to overlook efforts for the creation of publicly available datasets for new languages.
* If the availability of these new datasets has proven effective in fostering a better monitoring of misogynous hate speech in LDC or developing countries.
* If such monitoring has elsewhere proven to be an effective research, advocacy, and early warning tool.
* If current state-of-the-art technologies do not render likely the obsolescence of these datasets.

Then,

Accelerating the creation of datasets for hate speech and misogyny detection in underrepresented languages is a significant and impactful project.

To confirm the existence or absence of such datasets for specific Arabic dialects from North Africa, as well as Sahelian and Sub-Saharan languages, you can follow these steps:

**Research threads**

* Exploring hate speech dataset bias vs. real world performance
* Comparing LLM models over few datasets in Arabic and English, comparing performance between languages, analysis of performance trend, and variations in performance when changing pre-processing, augmentation and dataset split.

### Step-by-Step Approach

#### 1. \*\*Literature Review\*\*

- \*\*Research Papers\*\*: Start by searching for research papers in academic databases such as Google Scholar, IEEE Xplore, and ACM Digital Library. Use keywords like "hate speech detection", "misogyny detection", "Arabic dialects", "Sahelian languages", and "Sub-Saharan languages".

- \*\*Surveys and Reviews\*\*: Look for survey papers on hate speech detection or natural language processing (NLP) in African languages, as they often list existing datasets.

#### 2. \*\*Check Existing Repositories\*\*

- \*\*Hate Speech Data\*\*: Websites like [hatespeechdata.com](https://hatespeechdata.com) and [paperswithcode.com](https://paperswithcode.com/task/hate-speech-detection) often list datasets for hate speech detection. Browse these repositories for datasets in the languages of interest.

- \*\*General NLP Repositories\*\*: Check platforms like [Hugging Face Datasets](https://huggingface.co/datasets), [Kaggle](https://www.kaggle.com/datasets), and [NLP Hub](https://nlphub.io/).

#### 3. \*\*Language-Specific Resources\*\*

- \*\*Language Organizations\*\*: Contact organizations and research groups that focus on specific languages or regions. Examples include the Arabic Language Academy, African Language Technology Initiative, and universities with African studies departments.

- \*\*Regional Conferences\*\*: Look for proceedings from conferences like the Conference on African Languages and Linguistics (CALL) or the International Conference on Natural Language Processing (ICON).

#### 4. \*\*Social Media and Online Communities\*\*

- \*\*Social Media Analysis\*\*: Explore datasets from social media platforms where these languages are actively used. Platforms like Twitter, Facebook, and YouTube often contain large amounts of user-generated content.

- \*\*Online Forums\*\*: Engage with online communities on platforms like Reddit, Stack Exchange, or specialized forums where researchers discuss NLP and dataset creation.

#### 5. \*\*Collaborate with Local Institutions\*\*

- \*\*Universities and Research Institutes\*\*: Partner with universities and research institutes in regions where these languages are spoken. Local researchers might have unpublished datasets or ongoing projects related to hate speech detection.

- \*\*NGOs and Advocacy Groups\*\*: Work with non-governmental organizations and advocacy groups that focus on hate speech, human rights, or digital literacy in these regions.

#### 6. \*\*Direct Inquiries\*\*

- \*\*Emails and Surveys\*\*: Send inquiries to researchers, professors, and professionals who have published work related to your target languages. Use surveys or direct emails to ask about the existence of relevant datasets.

- \*\*Research Networks\*\*: Join research networks and mailing lists related to NLP, computational linguistics, and African studies to ask for information and collaborate.

### Steps to Create New Datasets

If you confirm the absence of datasets for the specific languages, you can proceed to create new ones. Here are the steps:

1. \*\*Data Collection\*\*:

- \*\*Sources\*\*: Collect data from social media, forums, news websites, and other online sources where the target languages are used.

- \*\*Ethical Considerations\*\*: Ensure ethical practices in data collection, respecting privacy and consent.

2. \*\*Annotation\*\*:

- \*\*Define Guidelines\*\*: Create clear guidelines for what constitutes hate speech and misogyny in the context of the target languages.

- \*\*Hire Annotators\*\*: Work with native speakers and linguists to accurately annotate the data. Train them on the annotation guidelines.

- \*\*Tools\*\*: Use annotation tools like Prodigy, Doccano, or custom-built interfaces.

3. \*\*Quality Control\*\*:

- \*\*Inter-Annotator Agreement\*\*: Measure the agreement between different annotators to ensure the consistency and reliability of the annotations.

- \*\*Review and Refine\*\*: Regularly review the annotations and refine the guidelines as needed.

4. \*\*Dataset Format\*\*:

- \*\*Standardization\*\*: Follow standard formats for dataset creation (e.g., CSV, JSON) and include metadata like source, date, and language/dialect information.

- \*\*Documentation\*\*: Provide thorough documentation describing the dataset creation process, guidelines, and potential applications.

5. \*\*Distribution\*\*:

- \*\*Open Access\*\*: Share the dataset on platforms like GitHub, Zenodo, or institutional repositories to make it accessible to the research community.

- \*\*Publication\*\*: Publish a paper describing the dataset and its creation process in relevant conferences or journals to reach a wider audience.

By following these steps, you can effectively identify the existence of datasets for the languages of interest and take the necessary steps to create and share new datasets if they do not already exist.

**### Recent Trends in Detecting Online Misogynistic Hate Speech**

The detection of online misogynistic hate speech is a growing field within natural language processing (NLP) and machine learning (ML). Recent studies and advancements have been made to address the challenges specific to this type of hate speech, focusing on both explicit and implicit forms.

\*\*1. \*\*Annotation and Taxonomy Development:\*\*

Creating detailed annotations and taxonomies for misogyny is essential. Recent work has involved expert-annotated datasets that categorize misogyny into specific types, such as anti-feminism, dehumanization, gendered slurs, and intersectional misogyny. These detailed annotations help in training more accurate models by providing clear definitions and examples of misogynistic language, which can vary significantly in expression and context【30†source】【31†source】.

\*\*2. \*\*Multi-target and Multi-task Learning:\*\*

Approaches that involve multi-target and multi-task learning have shown promise in improving the detection of misogynistic hate speech. By training models to recognize various forms of hate speech simultaneously, researchers have found that these models can transfer knowledge across different types of hate speech, improving detection accuracy. This method also helps in understanding the nuances between different hate speech categories and their specific targets, such as gender or ethnicity【29†source】.

\*\*3. \*\*Emotionally Informed Models:\*\*

Incorporating affective and emotional information into hate speech detection models is another emerging trend. Models that leverage resources like SenticNet and HurtLex, which provide affective and semantic insights, can better detect the emotional undertones of misogynistic statements. This approach helps in capturing the subtleties of language that convey misogyny, such as sarcasm or implicit bias【29†source】.

\*\*4. \*\*Challenges with Implicit Misogyny:\*\*

Detecting implicit misogyny remains a significant challenge. Implicit statements often lack overtly abusive language but still convey harmful sentiments. Recent research has focused on developing methods to identify these subtle forms of hate speech by using weakly supervised learning and leveraging large, unlabeled datasets. These methods aim to overcome the limitations of traditional supervised learning, which requires extensive annotated data【31†source】.

\*\*5. \*\*Performance Comparisons and Limitations:\*\*

While significant progress has been made, models for detecting misogynistic hate speech often face performance limitations, especially when compared to models for more commonly addressed forms of hate speech like racism. One reason is the complexity and variety of misogynistic language, which can be more context-dependent and culturally specific. Additionally, the availability of high-quality, annotated datasets for misogyny is less than for other types of hate speech, hindering model training and evaluation【30†source】【31†source】.

\*\*6. \*\*Practical Applications and Future Directions:\*\*

For practical applications, integrating these advanced detection models into social media monitoring tools and moderation systems is crucial. Future research directions include expanding the availability of diverse datasets, improving the interpretability of models, and developing more robust methods to handle the implicit and context-specific nature of misogynistic hate speech.

Overall, while the field is advancing, there is still a need for more comprehensive and nuanced approaches to detect and mitigate online misogynistic hate speech effectively.

NOTE

Si può comparare la performance di un Modello su più dataset o quella di più modelli su un dataset in particolare, **ma più frequentemente più modelli su più datasets.**