Dataset Card Template

1. Dataset Description

- **Homepage:** Add homepage URL here if available (unless it's a GitHub repository).
- **Repository:** If the dataset is hosted on GitHub or has a GitHub homepage, add the URL here.
- **Paper:** If the dataset was introduced by a paper or there was a paper written describing the dataset, add the URL here (landing page for Arxiv paper preferred).
- **Point of Contact:** If known, the name and email of at least one person the reader can contact for questions about the dataset.

1.1. Dataset Summary

Briefly summarize the dataset, its intended use, and the supported tasks. Give an overview of how and why the dataset was created. The summary should explicitly mention the languages present in the dataset (possibly in broad terms, e.g., *translations between several pairs of European languages*), and describe the domain, topic, or genre covered.

1.2. Languages

Provide a brief overview of the languages represented in the dataset. Describe relevant details about specifics of the language such as whether it is social media text, African, American English, etc.

2. Dataset Structure

2.1. Data Fields

List and describe the fields present in the dataset. Mention their data type, and whether they are used as input or output in any of the tasks the dataset currently supports. If the data has span indices, describe their attributes, such as whether they are at the character level or word level, whether they are contiguous or not, etc. If the datasets contain example IDs, state whether they have an inherent meaning, such as a mapping to other datasets or pointing to relationships between data points.

2.2. Data Splits

Describe and name the splits in the dataset if there are more than one.

Describe any criteria for splitting the data, if used. If there are differences between the splits (e.g. if the training annotations are machine-generated and the dev and test ones are created by humans, or if different numbers of annotators contributed to each example), describe them here.

Provide the sizes of each split. As appropriate, provide any descriptive statistics for the features, such as average length. For example:

	Train	Validation	Test
Input Sentences			
Average Sentence Length			

3. Dataset Creation

3.1. Curation Rationale

What need motivated the creation of this dataset? What are some of the reasons underlying the major choices involved in putting it together?

3.2. Source Data

This section describes the source data (e.g. news text and headlines, social media posts, translated sentences, source code, etc.).

3.2.1. Initial Data Collection and Normalization

If the dataset is new, describe the data collection process. Describe any criteria for data selection or filtering. List any keywords or search terms used. If possible, include runtime information for the collection process.

If data was collected from other pre-existing datasets, link to the source here.

If the data was modified or normalized after being collected (e.g. if the data is word-tokenized), describe the process and the tools used.

3.2.2. Who Are the Source Language Producers?

State whether the data was produced by humans or machine-generated. Describe the people or systems who originally created the data.

Describe the conditions under which the data was created (for example, if the producers were crowd workers, state what platform was used, or if the data was found, what website the data was found on). If compensation was provided, include that information here.

3.3. Annotations

If the dataset contains annotations which are not part of the initial dataset, describe them in the following sections.

3.3.1. Annotation Process

If applicable, describe the annotation process and any tools used, or state otherwise. Describe the amount of data annotated, if not all. Describe or reference annotation guidelines provided to the annotators. If available, provide inter-annotator statistics. Describe any annotation validation processes.

3.3.2. Who Are the Annotators?

If annotations were collected for the source data (such as class labels or syntactic parses), state whether the annotations were produced by humans or machine-generated.

Describe the people or systems who originally created the annotations and their selection criteria if applicable.

Describe the conditions under which the data was annotated (for example, if the annotators were crowd workers, state what platform was used, or if the data was found, what website the data was found on). If compensation was provided, include that information here.

3.4. Personal and Sensitive Information

State whether the dataset uses identity categories and, if so, how the information is used. Describe where this information comes from (i.e. self-reporting, collecting from profiles, inferring, etc.). State whether the data is linked to individuals and whether those individuals can be identified in the dataset, either directly or indirectly (i.e., in combination with other data).

State whether the dataset contains other data that might be considered sensitive (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history).

If efforts were made to anonymize the data, describe the anonymization process.

4. Considerations for Using the Data

4.1. Discussion of Biases

Provide descriptions of specific biases that are likely to be reflected in the data, and state whether any steps were taken to reduce their impact.

If analyses have been run quantifying these biases, please add brief summaries and links to the studies here.

4.2. Other Known Limitations

If studies of the datasets have outlined other limitations of the dataset, such as annotation artifacts, please outline and cite them here.

5. Additional Information

5.1. Dataset Curators

List the people involved in collecting the dataset and their affiliation(s). If funding information is known, include it here.

5.2. Licensing Information

Provide the license and link to the license webpage if available.

5.3. Citation Information

Provide the reference for the dataset. If the dataset has a DOI, please provide it here.

5.4. Contributions

Thanks to <<contributor>> for adding this dataset.