

# Data Papers. Eine kritische Bestandsaufnahme

Martin de la Iglesia, Caroline Jansky

Herzog August Bibliothek Wolfenbüttel / Zeitschrift für digitale Geisteswissenschaften

Einsatz von **CRedit** (Contributor Roles Taxonomy) für die Kennzeichnung der Verantwortlichkeiten der **Autor\*innen des Data Papers**. Rollen bei der Erstellung der Datenpublikation können abweichen. Mehrere Rollen pro Person und mehrere Personen pro Rolle möglich. Nicht alle CRediT-Rollen müssen genannt werden.

**Versionsangabe** und **Versionierungsdatum** des Data Papers

**Creative-Commons-Lizenz** der Zeitschrift (unabhängig von Lizenz der Forschungsdatenpublikation)

**Schlagwörter** (unter Nutzung kontrollierter Vokabulare) zu Thema, Methode sowie der adressierten Wissenschaftsdisziplin

**CRediT-Rollen** für die Kennzeichnung der Verantwortlichkeiten bei der **Erstellung der Datenpublikation**. Mehrere Rollen pro Person und mehrere Personen pro Rolle möglich. Nicht alle CRediT-Rollen müssen genannt werden.

Voraussetzungen für die Wahl des **Repositoriums**:

- Langzeitarchivierung
- persistente Adressierung
- keine Kosten
- keine weiteren Voraussetzungen für Zugriff (sofern rechtlich möglich)

Zeitschrift:

- Liste **empfohlener Repositorien** / **eigene Instanz** in einem den Voraussetzungen entsprechenden Repositorium
- Ausarbeitung von **Empfehlungen zur Verknüpfung / Veröffentlichung** von Data Paper und Gutachteninhalten im Repositorium

**Namensnennung Gutachter\*innen** optional, ggf. Identifikator wie ORCID nutzen; Grad der **Offenheit des Review-Verfahrens flexibel**, ausgehandelt von Redaktion, Autor\*innen und Gutachter\*innen

Darstellung der **Review-Ergebnisse** via Farbcodematrix, grob unterteilt in die **FAIR**-Bewertungskriterien Findability, Accessibility, Interoperability und Reusability

**Begutachtung** anhand einer kommentierten **Bewertungsmatrix**, basierend auf den **FAIR**-Prinzipien

**Veröffentlichung** der gesamten **Gutachteninhalte** (Bewertungsmatrix und Kommentare) in der Zeitschrift, im Data Paper verlinkt (und vice versa)

**Gutachten** ebenfalls **in der Datenpublikation** selbst verlinken oder dort eine Textversion der Gutachten veröffentlichen

**Zitierfähigkeit** bei ausführlichen Gutachten herstellen (eigener DOI, Zitierempfehlung, Seiten- oder Absatzzählung)

**Funktionen** von Data Papers:

- Erhöhte **Aufmerksamkeit** auf das Datenset, verbesserte **Auffindbarkeit**
- **Begutachtung** der Daten in standardisierten Verfahren (**Data Peer Review**)
- **Crediting** der an Datenpublikationen Beteiligten
- Strukturierte, ausführliche **Darstellung und Diskussion des Prozesses** der Datenerstellung, von Nachnutzungshorizonten und -limitationen der Datenpublikation sowie Entscheidungsprozessen

## A Curated Transformation of Sentence Dataset for Text Classification in Portuguese

Fulana de Tal<sup>a</sup>, João das Couves<sup>b</sup>

<sup>a</sup> Universidade Católica do Maracanã  
Contributions: [Writing – original draft](#), [Formal Analysis](#), [Supervision](#)  
fulanadetal@uni-maracana.edu

<sup>b</sup> Universidade Federal de Belém  
Contributions: [Writing – original draft](#), [Visualization](#)

DOI: [10.12345/678](#)

Published 12 November 2023

Last updated 29 February 2024

Version 2.0

License:

Keywords: [Machine learning](#); [Natural language processing](#); [Portuguese language](#); [Text classification](#); [Transformation of sentence](#)

Dataset: *PTSD: Portuguese Transformation of Sentence Dataset*

Contributors: João das Couves<sup>b</sup> (Universidade Federal de Belém; Contributions: [Conceptualization](#), [Data curation](#), [Formal Analysis](#)), Maria dos Anzóis<sup>c</sup> (Universidade Autónoma da Madeira; Contributions: [Investigation](#), [Methodology](#), [Resources](#), [Validation](#))

Version 2.0

Published 24 March 2023

Last updated 20 October 2023

License:

Repository: [Harvard Dataverse](#)

DOI: [10.54321/xyz](#)

Cite as: Das Couves, João, and Maria dos Anzóis. "PTSD: Portuguese Transformation of Sentence Dataset." 24 Feb. 2023. Version 2.0 from 20 Oct. 2023. Harvard Dataverse. DOI: [10.54321/xyz](#)

Data Review 1 – Caroline Jansky<sup>d</sup>

F A I R [Full Review](#)

Data Review 2 – Martin de la Iglesia<sup>e</sup>

F A I R [Full Review](#)

Refers to: De Tal, Fulana, and José da Silva. "Evaluation of Seven Text Classification Algorithms on a Portuguese Corpus." *European Review of Computational Linguistics* 22.2 (2023): 321–343. DOI: [10.98765/abc](#)

### 1. Background

Natural language processing (NLP) has seen significant advancements in recent years, driven by the availability of large-scale datasets. However, for languages other than English, resources are often limited, hindering progress in NLP research. Portuguese is one such language, where the availability of high-quality datasets is limited compared to English. To address this gap, we introduce the PTSD dataset, which focuses on transforming sentences in Portuguese to aid in various NLP tasks.

#### 1.1 Motivation

The motivation behind creating the PTSD dataset is to support research in Portuguese NLP. Portuguese is the most spoken language in the world, and it is essential to develop NLP models that can effectively handle Portuguese text. This dataset serves as a foundational resource for several tasks, including sentiment analysis, text classification, and machine translation.

Sentiment Analysis  
Text Classification  
Machine Translation  
Named Entity Recognition  
Text Summarization

By providing a diverse collection of transformed sentences, the PTSD dataset enables researchers to train and evaluate models for these tasks effectively.

### 2. Methods

#### 2.1 Data Collection

The PTSD dataset was created by collecting a diverse set of Portuguese sentences from various sources, including:

Online news articles  
Social media posts  
Books and literature  
Scientific papers

This wide range of sources ensures that the dataset reflects the diversity of the Portuguese language as used in different contexts.

#### 2.2 Data Transformation

To create the PTSD dataset, we applied several data transformation techniques to the collected sentences:

Tokenization: Each sentence was tokenized into words and punctuation marks.  
Lemmatization: We applied lemmatization to reduce words to their base forms, ensuring consistency and reducing data sparsity.  
Sentence Shuffling: Sentences were randomly shuffled to create variations and introduce diversity into the dataset.  
Synonym Replacement: Some words within sentences were replaced with synonyms, preserving sentence semantics while creating variations.  
Grammar and Structure Alteration: We introduced minor grammatical and structural changes to sentences to diversify the dataset further.  
The combination of these transformation techniques results in a rich and diverse dataset suitable for various NLP tasks.

### 3. Data Description

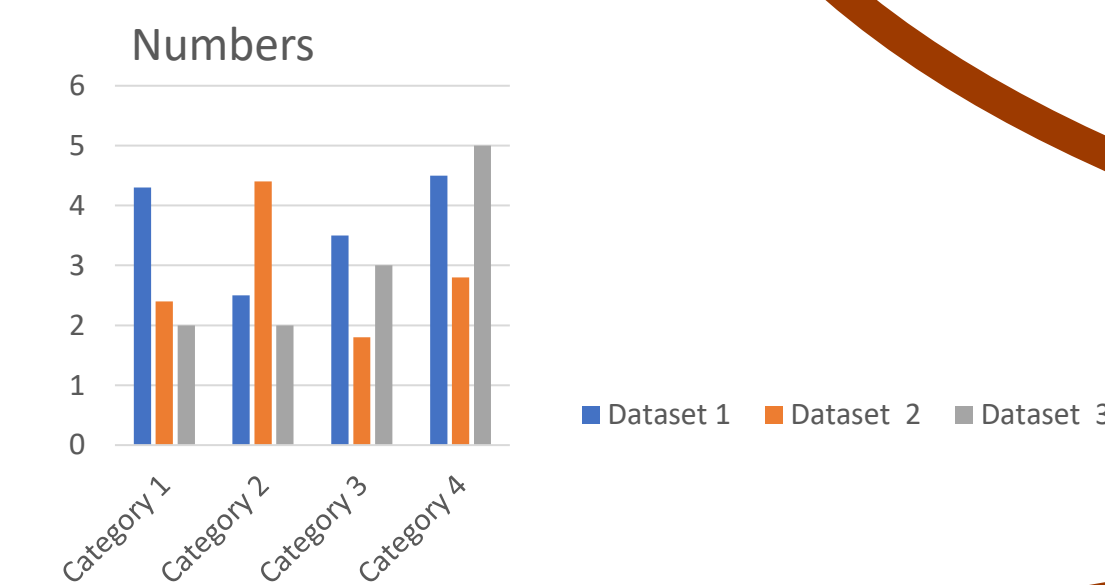
#### 3.1 Dataset Overview

The PTSD dataset consists of a total of 10,000 transformed Portuguese sentences. These sentences are divided into several categories, including news, social media, literature, and scientific domains, ensuring a broad representation of language usage.

#### 3.2 Data Format

Each sentence in the dataset is provided as a separate text file, with the following format:

ID: 12345  
Category: News  
Original Sentence: O governo anunciou novas políticas para a educação.  
Transformed Sentence: Anunciou o governo políticas novas para a educação.



#### 3.3 Data Statistics

Here are some key statistics about the PTSD dataset:

Total Sentences: 100,000  
Average Sentence Length: 15 words  
Categories: News, Social Media, Literature, Science  
Vocabulary Size: 30,000 unique words

### 4. Usage Notes

#### 4.1 Preprocessing

Researchers using the PTSD dataset should consider the following preprocessing steps:

Tokenization: Tokenize sentences into words and punctuation marks.  
Lemmatization: Apply lemmatization to reduce words to their base forms.  
Stopword Removal: Remove common stopwords to improve model efficiency.  
Data Split: Split the dataset into training, validation, and test sets for model development and evaluation.

#### 4.2 Model Training

The PTSD dataset can be used to train various NLP models, including neural networks, recurrent neural networks (RNNs), and transformer-based models like BERT and GPT. Researchers are encouraged to experiment with different architectures and hyperparameters to achieve optimal results for their specific task.

#### 4.3 Evaluation Metrics

For tasks such as sentiment analysis and text classification, common evaluation metrics such as accuracy, precision, recall, and F1-score can be used. Researchers should choose appropriate metrics based on their specific NLP task.

### References

- [1] McEnery, Tony, and Andrew Hardie. *Corpus linguistics: Method, theory and practice*. Cambridge University Press, 2011.
- [2] Schmit, Cristina. "Cross-linguistic variation and the present perfect: The case of Portuguese." *Natural language & linguistic theory* 19.2 (2001): 403–453.

**Autor\*innen** des Data Papers getrennt von den Beteiligten an der Datenpublikation aufgeführt

**Identifikation** durch Angabe der institutionellen Affiliationen sowie ORCID

**DOI** als **Standard-PID** für Zeitschriftenbeiträge

DOI des Data Papers nicht identisch mit DOI bzw. PID der Datenpublikation

**Titel** der Datenpublikation im Repositorium

An der **Datenpublikation beteiligte Personen**, nicht zwingend identisch mit den Autor\*innen des Data Papers. Benennung Hauptverantwortliche, Nennung weiterer Beteiligter ist möglich.

**Version** der **Datenpublikation**, auf die sich das Data Paper bezieht

**Datum der Erstveröffentlichung** und **Versionierungsdatum** der Datenpublikation

Veröffentlichung der Daten unter einer **Creative-Commons-Lizenz** wünschenswert, aber nicht zwingende Voraussetzung

Data Papers zu eingeschränkt zugänglichen Daten nicht rigoros ausschließen

Voraussetzungen hinsichtlich **Zugänglichkeit**: Daten müssen

a) Redaktion und Gutachter\*innen zur Verfügung stehen, und

b) für konkrete Forschungsvorhaben verfügbar gemacht werden können

**Zitation** des Data Papers nur in Ausnahmefällen notwendig, stattdessen **Zitierempfehlung** für die Datenpublikation

Angabe von **Forschungspublikationen** im Zusammenhang mit der Forschungsdatenpublikation

Festgelegte **Struktur** des Texts:

1. Grundlegende Zielstellung / Projektzusammenhang / Datenquellen
2. Methoden der Datenerhebung, -bereinigung und -aufbereitung
3. Beschreibung der Struktur der Datenpublikation / technische Spezifikationen der Daten
4. Publierte Forschung auf bzw. zu den Daten / potenzielle Nutzungshorizonte / Limitationen der Nutzung / Related Works

**Schwerpunktsetzung** durch Unterkapitel und unterschiedliche Ausführlichkeit

**Voraussetzungen** für die Publikation eines Data Papers:

Editorial Pre-Review hinsichtlich

- inhaltlicher Ausrichtung
- struktureller Anforderungen
- Interesses der Community an Rezeption und Nachnutzung der Daten

Poster im Rahmen der FORGE23 – Anything goes? Forschungsdaten in den Geisteswissenschaften kritisch betrachtet. Tübingen, 4. bis 6. Oktober 2023.