Hauptsache ›Open‹?

Data Papers für die (digitalen) Geisteswissenschaften

Caroline Jansky / Redaktionsleitung der Zeitschrift für digitale Geisteswissenschaften
Herzog August Bibliothek Wolfenbüttel / Forschungsverbund Marbach Weimar Wolfenbüttel

HERZOG AUGUST BIBLIOTHEK
DHd digital humanities im deutschsprachigen raum
ZfdG
MARBACH WEIMAR WOLFENBÜTTEL FORSCHUNGSVERBUND
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

Gliederung

Zeitschrift für digitale Geisteswissenschaften

- Inhalte
- Struktur
- Formate

Data Papers für die (digitalen) Geisteswissenschaften

- Brückenformat: Data Paper → Datenpublikation
- Peer Review: Dimensionen der Openness

Zusammenfassung

- Funktionen von Data Papers
- Openness als Ressourcenfrage

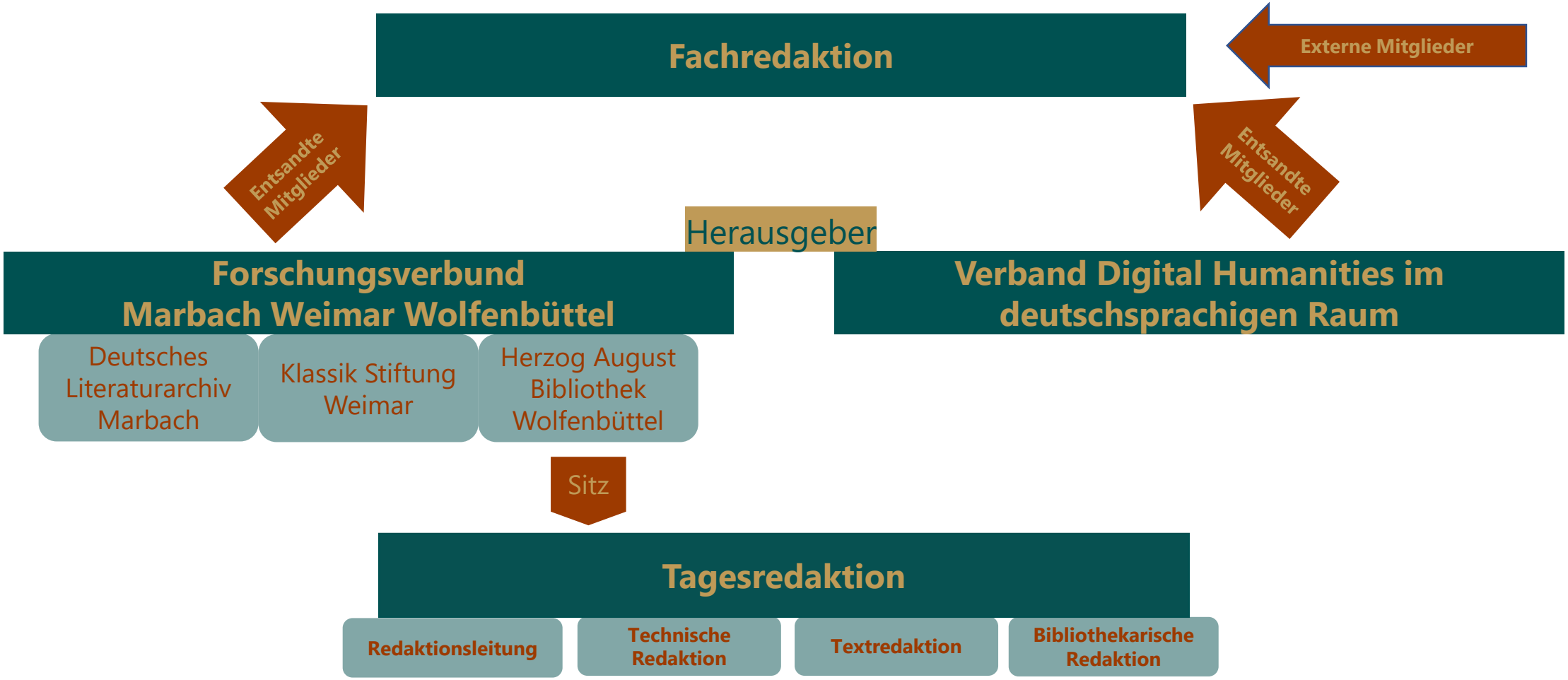
Zeitschrift für digitale Geisteswissenschaften

ZfdG

Publiziert seit 2015 Beiträge (digital, Diamond-Open-Access-Modell):

- an der Schnittstelle von Informatik / Informationswissenschaft und Geisteswissenschaften
- zu sozial- und gesellschaftswissenschaftlichen Fragestellungen, die sich digitaler Methoden bedienen
- zu theoretischen Reflexionen und den ›Humanities of the Digital‹

Zeitschrift für digitale Geisteswissenschaften

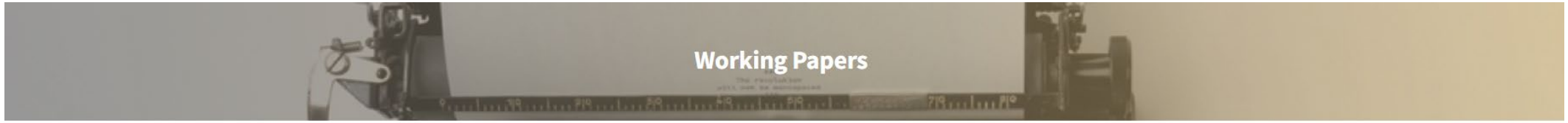


Zeitschrift für digitale Geisteswissenschaften

ZfdG

- *Jahreshefte*: Fachartikel, Projektvorstellungen
- *Sonderbände* (in externer Herausgeberschaft): Fachartikel [Long Papers], Projektvorstellungen, [Data Papers]
- *Working Papers*: eigene Publikationsreihe, Diskussions- und Grundlagenpapiere

Working Papers



Die Reihe »Working Papers der ZfdG« umfasst Diskussions- und Arbeitspapiere aus allen Bereichen der Digital Humanities: Positionspapiere, Best Practices, Dokumentationen von Prozessen oder Modellen und Erfahrungsberichte finden hier einen Publikationsort.

DOI: 10.17175/working-papers

Im Public-Peer-Review-Verfahren werden die Working Papers durch die Community mittels Hypothes.is kommentiert, ergänzt und kritisch begutachtet. Nach der Überarbeitung durch die Autor*innen werden die finalen Versionen der Working Papers statisch publiziert und beide Versionen miteinander verlinkt, so dass die Änderungen nachvollzogen werden können.

Referenzrahmen für eigenständige digitale Wissenschaftskommunikation durch Forschende

Claudia Frick / Melanie Seltmann

Begriffe der Digital Humanities. Ein diskursives Glossar

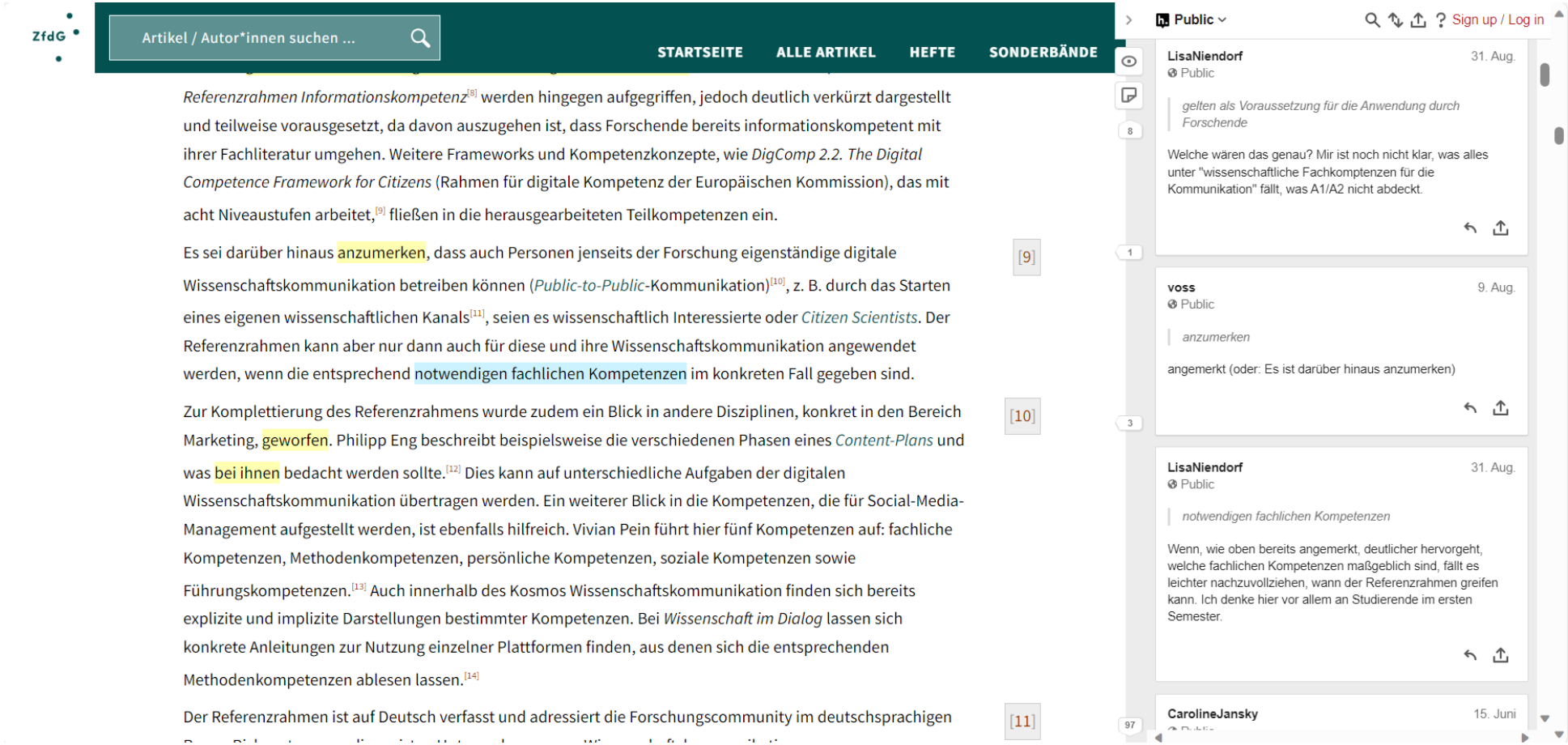
AG Theorie des Verbandes Digital Humanities im deutschsprachigen Raum e. V.

Öffentliche Begutachtung bis zum 01.09.2023

Digitales Publizieren in den Geisteswissenschaften: Begriffe, Standards, Empfehlungen

AG Digitales Publizieren des Verbandes Digital Humanities im deutschsprachigen Raum e. V.

Working Papers: Open Public Peer Review



Working Papers: Versionierung

166	Bedarf, das Working Paper zu überarbeiten. Nicht zuletzt machte die	156	sie?
167	mittlerweile festzustellende Etablierung des digitalen Publizierens im		
168	Wissenschaftsbetrieb es auch notwendig, andere Schwerpunkte zu setzen. In		
169	der hier vorliegenden zweiten Version wird daher stärker auf aktuelle		
170	Entwicklungen und Diskurse denn auf grundlegende Informationen und		
171	Empfehlungen fokussiert. Daher wird bewusst nur am Rand auf die Formate und		
172	Strukturen des traditionellen Publizierens Bezug genommen. Neu hinzugekommen		
173	ist das Kapitel zu Publikationsinfrastrukturen. Alle anderen Kapitel wurden		
174	grundlegend überarbeitet.[5]		
175		157	
176	[3]Die Literatur zu dem Working Paper kann über eine Zotero Library eingesehen werden.[6] Informationen und Neuigkeiten zur AG werden über die Mailingliste der AG kommuniziert, auf die sich jede*r	158	[5]Digitale wissenschaftliche Publikationen zeichnen sich durch die Möglichkeiten des
177	Interessierte gerne eintragen kann. Das vorliegende Paper wird in einem	159	Mediums, wie Maschinenlesbarkeit, Multimedialität, Veränderbarkeit, leichte Kopierbarkeit,
178	offenen Begutachtungsverfahren durch die Community begutachtet. Die	160	Vernetzbarkeit, etc. aus und erweitern methodisch die Verfahren wissenschaftlicher
179	Begutachtung erfolgt über die Zeitschrift für digitale Geisteswissenschaft	161	Ergebnissicherung. Die neuen digitalen Möglichkeiten haben den Begriff der Publikation
180	(ZfdG) unter Verwendung des Open-Source-Tools Hypothesis. Die im Rahmen des	162	erweitert, der die ehemals etablierten Publikationswege über das klassische Verlagsmodell
181	Public-Review-Verfahrens angemerkten Verbesserungsvorschläge und	163	nur als eine von mehreren Optionen der Veröffentlichung versteht. Der folgende Abriss
182	Diskussionspunkte werden in eine überarbeitete Version einfließen, die	164	geht daher von einem weiten Publikationsbegriff aus. Dazu zählen ohne Anspruch auf
183	Version mit den Kommentaren bleibt ebenso veröffentlicht.	165	Vollständigkeit:
184		166	
185		167	Digitale Texte in traditionellen Formen (Monografien, Sammelbände, Aufsätze, Rezensionen,
186	1. Was sind digitale wissenschaftliche Publikationen und welche	168	Editionen, Kommentare)
187	Möglichkeiten bieten sie?		
188			

Data Papers als Brückenformat

Forschungsbeiträge in Zeitschriften (peer reviewed)

FAIRe Forschungsdatenpublikation (qualitätsgesichert, standardisiert, kontextualisiert)

Data Papers: Bestandsaufnahme

- Research Data Journal for the Humanities and Social Sciences (RDJ)
- Journal of Open Humanities Data (JOHD)
- Journal of Data Mining and Digital Humanities
- NECSUS
- Journal of Digital History
- Digital Humanities Quarterly
- International Journal for Digital Art History
- Scientific Data

Skizze eines „idealen“ Data Papers für die (digitalen) Geisteswissenschaften

Umsetzung in der ZfdG im Laufe des Jahres 2024

A Curated Transformation of Sentence Dataset for Text Classification in Portuguese

Fulana de Tal[a], João das Couves[b]

[a] Universidade Católica do Maracanã
Contributions: Writing – original draft, Formal Analysis, Supervision
fulanadetal@uni-maracana.edu

[b] 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 (Universidade Federal de Belém; Contributions: Conceptualization, Data curation, Formal Analysis), Maria dos Anzóis (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
F A I R Full Review

Data Review 2 – Martin de la Iglesia
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 23.2 (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 and diverse 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 one of the most widely spoken languages 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 NLP tasks, including but not limited to:

- 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.

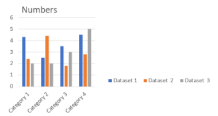
The combination of these transformation techniques results in a rich and diverse dataset suitable for various NLP tasks.

3. Data Description

3.1 Dataset Overview
The PTSD dataset consists of a total of 100,000 transformed Portuguese sentences. These sentences are divided into several categories, including news, social media, literature, and scientific domains, ensuring a broad representation of language use.

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: 50,000 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.

Poster: Martin de la Iglesia / Caroline Jansky: Data Papers - Eine kritische Bestandsaufnahme. FORGE 2023 - Anything Goes?! Forschungsdaten in den Geisteswissenschaften - kritisch betrachtet (FORGE2023), Tübingen, Deutschland. Zenodo. DOI: 10.5281/zenodo.8392501

A Curated Transformation of Sentence Dataset for Text Classification in Portuguese

Fulana de Tal[a], João das Couves[b]

[a] Universidade Católica do Maracanã
Contributions: Writing – original draft, Formal Analysis, Supervision
fulanadetal@uni-maracana.edu

[b] 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: CC BY
Keywords: Machine learning; Natural language processing; Portuguese language; Text classification; Transformation of sentence

Dataset: *PTSD: Portuguese Transformation of Sentence Dataset*
Contributors: João das Couves (Universidade Federal de Belém; Contributions: Conceptualization, Data curation, Formal Analysis), Maria dos Anzóis (Universidade Autónoma da Madeira; Contributions: Investigation, Methodology, Resources, Validation)
Version 2.0
Published 24 March 2023
Last updated 20 October 2023
License: PUBLIC DOMAIN
Repository: Harvard Dataverse
DOI: 10.54321/xyz
Cite as: Das Couves, João, and Maria dos Anzóis. "PTSD: Portuguese Transformation of Sentenc
Dataset." 24 Feb. 2023. Version 2.0 from 20 Oct. 2023. Harvard Dataverse. DOI: 10.54321/xyz

Data Review 1 – Caroline Jansky
F A I R Full Review

Versionsangabe und **Versionierungsdatum** des Data Papers

Version der Datenpublikation, auf die sich das Data Paper bezieht

Datum der Erstveröffentlichung und **Versionierungsdatum** der Datenpublikation

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

DOI als **Standard-PID** für Zeitschriftenbeiträge

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

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

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

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

Gutachten ebenfalls **in der Datenpublikation** veröffentlichen oder verlinken

classification; Transformation of sentence

Dataset: *PTSD: Portuguese Transformation of Sentence Dataset*
Contributors: João das Couves (Universidade Federal de Belém; Contributions: Conceptualization, Data curation, Formal Analysis), Maria dos Anzóis (Universidade Autónoma da Madeira; Contributions: Investigation, Methodology, Resources, Validation)
Version 2.0
Published 24 March 2023
Last updated 20 October 2023
License: PUBLIC DOMAIN
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
F A I R Full Review

Data Review 2 – Martin de la Iglesia
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 23.2 (2023): 321–343. DOI: 10.98765/abc

Darstellung der **Review-Ergebnisse** via Farbcode

1. Background

Natural language processing (NLP) has seen significant advancements in recent years, driven by the availability of large and diverse 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 one of the most widely spoken languages 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 NLP tasks, including but not limited to:

- 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

Peer Review: Dimensionen der Openness

- *Open identities* (Beteiligte einander namentlich bekannt)
- *Open reports* (Gutachten veröffentlicht)
- *Open participation* (Beteiligung der Community)
- *Open pre-review manuscripts* (Veröffentlichung vor Review)
- *Open interaction* (alle Beteiligten im Austausch)
- *Open final-version commenting* (Kommentierung finale Version)
- *Open platforms* (Review von unabhängigem Anbieter)[1]

[1] Tony Ross-Hellauer: What Is Open Peer Review? A Systematic Review. In: F1000 Research 2017, 6:588. 27.04.2017. Version 2.0 vom 31.08.2017, S. 7. DOI: 10.12688/f1000research.11369.2; vgl. auch Yuliya Fadeeva: Qualitative Sprünge in der Qualitätssicherung? Potenziale digitaler Open-Peer-Review-Formate. In: Fabrikation von Erkenntnis – Experimente in den Digital Humanities. Hg. von Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels Walkowski, Joëlle Weis, Ulrike Wuttke. Wolfenbüttel 2021. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 5). 01.09.2021. Version 2.0 vom 21.03.2023. HTML / XML / PDF. DOI: 10.17175/sb005_002_v2

Peer Review: Dimensionen der Openness

- *Open identities* (Beteiligte einander namentlich bekannt)
- *Open reports* (Gutachten veröffentlicht)
- *Open participation* (Beteiligung der Community)
- *Open pre-review manuscripts* (Veröffentlichung vor Review)
- *Open interaction* (alle Beteiligten im Austausch)
- *Open final-version commenting* (Kommentierung finale Version)
- *Open platforms* (Review von unabhängigem Anbieter)[1]

Im Open Public Peer Review umgesetzt | teilweise im Open [Public] Peer Review umgesetzt | vollständig umgesetzt im Open [Public] Peer Review

[1] Tony Ross-Hellauer: What Is Open Peer Review? A Systematic Review. In: F1000 Research 2017, 6:588. 27.04.2017. Version 2.0 vom 31.08.2017, S. 7. DOI: 10.12688/f1000research.11369.2; vgl. auch Yuliya Fadeeva: Qualitative Sprünge in der Qualitätssicherung? Potenziale digitaler Open-Peer-Review-Formate. In: Fabrikation von Erkenntnis – Experimente in den Digital Humanities. Hg. von Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels Walkowski, Joëlle Weis, Ulrike Wuttke. Wolfenbüttel 2021. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 5). 01.09.2021. Version 2.0 vom 21.03.2023. HTML / XML / PDF. DOI: 10.17175/sb005_002_v2

Natural language processing (NLP) has seen significant advancements in recent years, driven by the availability of large and diverse 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 one of the most widely spoken languages 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 NLP tasks, including but not limited to:

- 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

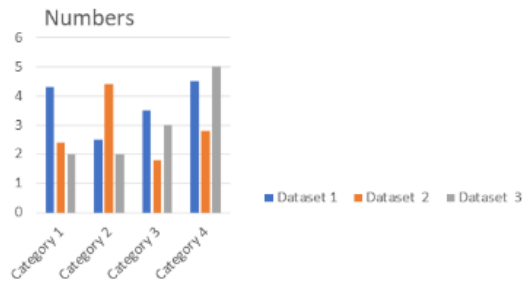
3.1 Dataset Overview

The PTSD dataset consists of a total of 100,000 transformed Portuguese sentences. These sentences are divided into several categories, including news, social media, literature, and scientific domains, ensuring a broad representation of language use.

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

Festgelegte **Struktur** des Texts:

1. Grundlegende Zielstellung
2. Methoden der Datenerhebung
3. Beschreibung der Datenpublikation / technische Spezifikationen der Daten
4. potenzielle Nutzungshorizonte / Limitationen / Related Works

Schwerpunktsetzung durch Unterkapitel und unterschiedliche Ausführlichkeit

classification; Transformation of sentence

Dataset: *PTSD: Portuguese Transformation of Sentence Dataset*

Contributors: João das Couves (Universidade Federal de Belém; Contributions: Conceptualization, Data curation, Formal Analysis), Maria dos Anzóis (Universidade Autónoma da Madeira; Contributions: Investigation, Methodology, Resources, Validation)

Version 2.0
Published 24 March 2023
Last updated 20 October 2023
License: PUBLIC DOMAIN
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

F	A	I	R	
●	●	●	●	Full Review

Data Review 2 – Martin de la Iglesia

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 23.2 (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 and diverse 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 one of the most widely spoken languages 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 NLP tasks, including but not limited to:

- 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

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

Veröffentlichung der gesamten **Gutachteninhalte**

Gutachten ebenfalls **in der Datenpublikation** veröffentlichen oder verlinken

Zitierfähigkeit der Gutachten (eigener DOI, Zitierempfehlung, Seiten- oder Absatzzählung)

Namensnennung Gutachter*innen optional

Grad der **Offenheit des Review-Verfahrens flexibel**, ausgehandelt von Redaktion, Autor*innen und Gutachter*innen

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

Peer Review: Dimensionen der Openness

- *Open identities* (Beteiligte einander namentlich bekannt)
- *Open reports* (Gutachten veröffentlicht)
- *Open participation* (Beteiligung der Community)
- *Open pre-review manuscripts* (Veröffentlichung vor Review)
- *Open interaction* (alle Beteiligten im Austausch)
- *Open final-version commenting* (Kommentierung finale Version)
- *Open platforms* (Review von unabhängigem Anbieter)[1]

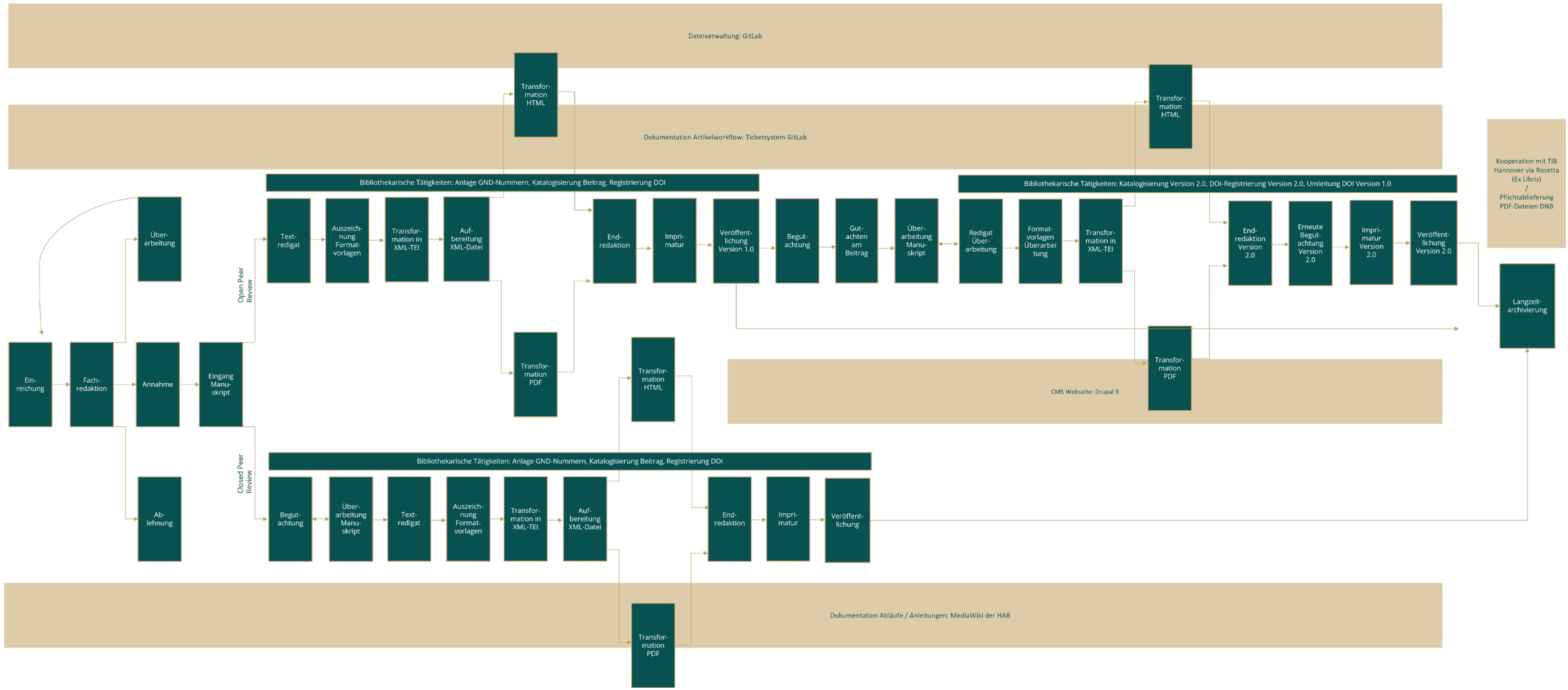
Im Open Public Peer Review umgesetzt | teilweise im Open [Public] Peer Review umgesetzt | vollständig umgesetzt im Open [Public] Peer Review

[1] Tony Ross-Hellauer: What Is Open Peer Review? A Systematic Review. In: F1000 Research 2017, 6:588. 27.04.2017. Version 2.0 vom 31.08.2017, S. 7. DOI: 10.12688/f1000research.11369.2; vgl. auch Yuliya Fadeeva: Qualitative Sprünge in der Qualitätssicherung? Potenziale digitaler Open-Peer-Review-Formate. In: Fabrikation von Erkenntnis – Experimente in den Digital Humanities. Hg. von Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels Walkowski, Joëlle Weis, Ulrike Wuttke. Wolfenbüttel 2021. (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 5). 01.09.2021. Version 2.0 vom 21.03.2023. HTML / XML / PDF. DOI: 10.17175/sb005_002_v2

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

Openness als Ressourcenfrage



Dank für die Aufmerksamkeit!

Kontakt zur Redaktion: zfdg@mww-forschung.de
Kontakt zur Referentin: jansky@hab.de