

Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)

St. Julian's, Malta
21 March 2024

ISBN: 978-1-7138-9448-3

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2024) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006

Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>Handling Name Errors of a BERT-Based De-Identification System: Insights from Stratified Sampling and Markov-based Pseudonymization</i>	
Dalton Simancek and VG Vinod Vydiswaran	1
<i>Assessing Authenticity and Anonymity of Synthetic User-generated Content in the Medical Domain</i>	
Tomohiro Nishiyama, Lisa Raithel, Roland Roller, Pierre Zweigenbaum and Eiji Aramaki	8
<i>Automatic Detection and Labelling of Personal Data in Case Reports from the ECHR in Spanish: Evaluation of Two Different Annotation Approaches</i>	
Maria Sierro, Begoña Altuna and Itziar Gonzalez-Dios	18
<i>PSILENCE: A Pseudonymization Tool for International Law</i>	
Luis Adrián Cabrera-Diego and Akshita Gheewala	25
<i>Deidentifying a Norwegian Clinical Corpus - an Effort to Create a Privacy-preserving Norwegian Large Clinical Language Model</i>	
Phuong Ngo, Miguel Tejedor, Therese Olsen Svenning, Taridzo Chomutare, Andrius Budrionis and Hercules Dalianis	37
<i>Extending Off-the-shelf NER Systems to Personal Information Detection in Dialogues with a Virtual Agent: Findings from a Real-Life Use Case</i>	
Mario Mina, Carlos Rodríguez, Aitor Gonzalez-Agirre and Marta Villegas	44
<i>Detecting Personal Identifiable Information in Swedish Learner Essays</i>	
Maria Irena Szawerna, Simon Dobnik, Ricardo Muñoz Sánchez, Therese Lindström Tiedemann and Elena Volodina	54
<i>Data Anonymization for Privacy-Preserving Large Language Model Fine-Tuning on Call Transcripts</i>	
Shayna Gardiner, Tania Habib, Kevin Humphreys, Masha Azizi, Frederic Mailhot, Anne Paling, Preston Thomas and Nathan Zhang	64
<i>When Is a Name Sensitive? Eponyms in Clinical Text and Implications for De-Identification</i>	
Thomas Vakili, Tyr Hullmann, Aron Henriksson and Hercules Dalianis	76
<i>Did the Names I Used within My Essay Affect My Score? Diagnosing Name Biases in Automated Essay Scoring</i>	
Ricardo Muñoz Sánchez, Simon Dobnik, Maria Irena Szawerna, Therese Lindström Tiedemann and Elena Volodina	81