


## Series Editor

Janusz Kacprzyk , *Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*

## Advisory Editors

Fernando Gomide, *Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil*

Okyay Kaynak, *Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Türkiye*

Derong Liu, *Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA*

*Institute of Automation, Chinese Academy of Sciences, Beijing, China*

Witold Pedrycz, *Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada*

*Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*

Marios M. Polycarpou, *Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus*

Imre J. Rudas, *Óbuda University, Budapest, Hungary*

Jun Wang, *Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong*

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the worldwide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose ([aninda.bose@springer.com](mailto:aninda.bose@springer.com)).

Álvaro Rocha · Hojjat Adeli ·  
Gintautas Dzemyda · Fernando Moreira ·  
Aneta Poniszewska-Marańda  
Editors

# Good Practices and New Perspectives in Information Systems and Technologies

WorldCIST 2024, Volume 6

 Springer

*Editors*

Álvaro Rocha  
ISEG  
Universidade de Lisboa  
Lisbon, Portugal

Hojjat Adeli  
College of Engineering  
The Ohio State University  
Columbus, OH, USA

Gintautas Dzemyda  
Institute of Data Science and Digital  
Technologies  
Vilnius University  
Vilnius, Lithuania

Fernando Moreira  
DCT  
Universidade Portucalense  
Porto, Portugal

Aneta Poniszewska-Marañda  
Institute of Information Technology  
Lodz University of Technology  
Łódź, Poland

ISSN 2367-3370 ISSN 2367-3389 (electronic)  
Lecture Notes in Networks and Systems  
ISBN 978-3-031-60327-3 ISBN 978-3-031-60328-0 (eBook)  
<https://doi.org/10.1007/978-3-031-60328-0>

© The Editor(s) (if applicable) and The Author(s), under exclusive license  
to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

## Preface

This book contains a selection of papers accepted for presentation and discussion at the 2024 World Conference on Information Systems and Technologies (WorldCIST'24). This conference had the scientific support of the Lodz University of Technology, Information and Technology Management Association (ITMA), IEEE Systems, Man, and Cybernetics Society (IEEE SMC), Iberian Association for Information Systems and Technologies (AISTI), and Global Institute for IT Management (GIIM). It took place in Lodz city, Poland, 26–28 March 2024.

The World Conference on Information Systems and Technologies (WorldCIST) is a global forum for researchers and practitioners to present and discuss recent results and innovations, current trends, professional experiences, and challenges of modern Information Systems and Technologies research, technological development, and applications. One of its main aims is to strengthen the drive toward a holistic symbiosis between academy, society, and industry. WorldCIST'23 is built on the successes of: WorldCIST'13 held at Olhão, Algarve, Portugal; WorldCIST'14 held at Funchal, Madeira, Portugal; WorldCIST'15 held at São Miguel, Azores, Portugal; WorldCIST'16 held at Recife, Pernambuco, Brazil; WorldCIST'17 held at Porto Santo, Madeira, Portugal; WorldCIST'18 held at Naples, Italy; WorldCIST'19 held at La Toja, Spain; WorldCIST'20 held at Budva, Montenegro; WorldCIST'21 held at Terceira Island, Portugal; WorldCIST'22 held at Budva, Montenegro; and WorldCIST'23, which took place at Pisa, Italy.

The Program Committee of WorldCIST'24 was composed of a multidisciplinary group of 328 experts and those who are intimately concerned with Information Systems and Technologies. They have had the responsibility for evaluating, in a 'blind review' process, the papers received for each of the main themes proposed for the conference: A) Information and Knowledge Management; B) Organizational Models and Information Systems; C) Software and Systems Modeling; D) Software Systems, Architectures, Applications and Tools; E) Multimedia Systems and Applications; F) Computer Networks, Mobility and Pervasive Systems; G) Intelligent and Decision Support Systems; H) Big Data Analytics and Applications; I) Human-Computer Interaction; J) Ethics, Computers & Security; K) Health Informatics; L) Information Technologies in Education; M) Information Technologies in Radiocommunications; and N) Technologies for Biomedical Applications.

The conference also included workshop sessions taking place in parallel with the conference ones. Workshop sessions covered themes such as: ICT for Auditing & Accounting; Open Learning and Inclusive Education Through Information and Communication Technology; Digital Marketing and Communication, Technologies, and Applications; Advances in Deep Learning Methods and Evolutionary Computing for Health Care; Data Mining and Machine Learning in Smart Cities: The role of the technologies in the research of the migrations; Artificial Intelligence Models and Artifacts for Business Intelligence Applications; AI in Education; Environmental data analytics; Forest-Inspired

Computational Intelligence Methods and Applications; Railway Operations, Modeling and Safety; Technology Management in the Electrical Generation Industry: Capacity Building through Knowledge, Resources and Networks; Data Privacy and Protection in Modern Technologies; Strategies and Challenges in Modern NLP: From Argumentation to Ethical Deployment; and Enabling Software Engineering Practices Via Last Development Trends.

WorldCIST'24 and its workshops received about 400 contributions from 47 countries around the world. The papers accepted for oral presentation and discussion at the conference are published by Springer (this book) in four volumes and will be submitted for indexing by WoS, Scopus, EI-Compendex, DBLP, and/or Google Scholar, among others. Extended versions of selected best papers will be published in special or regular issues of leading and relevant journals, mainly JCR/SCI/SSCI and Scopus/EI-Compendex indexed journals.

We acknowledge all of those that contributed to the staging of WorldCIST'24 (authors, committees, workshop organizers, and sponsors). We deeply appreciate their involvement and support that was crucial for the success of WorldCIST'24.

March 2024

Álvaro Rocha  
Hojjat Adeli  
Gintautas Dzemyda  
Fernando Moreira  
Aneta Poniszewska-Marańda



Chris Kimble	KEDGE Business School & MRM, UM2, Montpellier, France
Damian Niwiński	University of Warsaw, Poland
Eugene Spafford	Purdue University, USA
Florin Gheorghe Filip	Romanian Academy, Romania
Janusz Kacprzyk	Polish Academy of Sciences, Poland
João Tavares	University of Porto, Portugal
Jon Hall	The Open University, UK
John MacIntyre	University of Sunderland, UK
Karl Stroetmann	Empirica Communication & Technology Research, Germany
Marjan Mernik	University of Maribor, Slovenia
Miguel-Angel Sicilia	University of Alcalá, Spain
Mirjana Ivanovic	University of Novi Sad, Serbia
Paulo Novais	University of Minho, Portugal
Sami Habib	Kuwait University, Kuwait
Wim Van Grembergen	University of Antwerp, Belgium

### **Program Committee Co-chairs**

Adam Wojciechowski	Lodz University of Technology, Poland
Aneta Poniszewska-Marańda	Lodz University of Technology, Poland

### **Program Committee**

Abderrahmane Ez-zahout	Mohammed V University, Morocco
Adriana Peña Pérez Negrón	Universidad de Guadalajara, Mexico
Adriani Besimi	South East European University, North Macedonia
Agostinho Sousa Pinto	Polytechnic of Porto, Portugal
Ahmed El Oualkadi	Abdelmalek Essaadi University, Morocco
Akex Rabasa	University Miguel Hernandez, Spain
Alanio de Lima	UFC, Brazil
Alba Córdoba-Cabús	University of Malaga, Spain
Alberto Freitas	FMUP, University of Porto, Portugal
Aleksandra Labus	University of Belgrade, Serbia
Alessio De Santo	HE-ARC, Switzerland
Alexandru Vulpe	University Politehnica of Bucharest, Romania
Ali Idri	ENSIAS, University Mohamed V, Morocco
Alicia García-Holgado	University of Salamanca, Spain



Almir Souza Silva Neto	IFMA, Brazil
Álvaro López-Martín	University of Malaga, Spain
Amélia Badica	Universiti of Craiova, Romania
Amélia Cristina Ferreira Silva	Polytechnic of Porto, Portugal
Amit Shelef	Sapir Academic College, Israel
Ana Carla Amaro	Universidade de Aveiro, Portugal
Ana Dinis	Polytechnic of Cávado and Ave, Portugal
Ana Isabel Martins	University of Aveiro, Portugal
Anabela Gomes	University of Coimbra, Portugal
Anacleto Correia	CINAV, Portugal
Andrew Brosnan	University College Cork, Ireland
Andjela Draganic	University of Montenegro, Montenegro
Aneta Polewko-Klim	University of Białystok, Institute of Informatics, Poland
Aneta Poniszewska-Maranda	Lodz University of Technology, Poland
Angeles Quezada	Instituto Tecnológico de Tijuana, Mexico
Anis Tissaoui	University of Jendouba, Tunisia
Ankur Singh Bist	KIET, India
Ann Svensson	University West, Sweden
Anna Gawrońska	Poznański Instytut Technologiczny, Poland
Antoni Oliver	University of the Balearic Islands, Spain
Antonio Jiménez-Martín	Universidad Politécnica de Madrid, Spain
Aroon Abbu	Bell and Howell, USA
Arslan Enikeev	Kazan Federal University, Russia
Beatriz Berrios Aguayo	University of Jaen, Spain
Benedita Malheiro	Polytechnic of Porto, ISEP, Portugal
Bertil Marques	Polytechnic of Porto, ISEP, Portugal
Boris Shishkov	ULSIT/IMI - BAS/IICREST, Bulgaria
Borja Bordel	Universidad Politécnica de Madrid, Spain
Branko Perisic	Faculty of Technical Sciences, Serbia
Bruno F. Gonçalves	Polytechnic of Bragança, Portugal
Carla Pinto	Polytechnic of Porto, ISEP, Portugal
Carlos Balsa	Polytechnic of Bragança, Portugal
Carlos Rompante Cunha	Polytechnic of Bragança, Portugal
Catarina Reis	Polytechnic of Leiria, Portugal
Célio Gonçalo Marques	Polytechnic of Tomar, Portugal
Cengiz Acarturk	Middle East Technical University, Turkey
Cesar Collazos	Universidad del Cauca, Colombia
Cristina Gois	Polytechnic University of Coimbra, Portugal
Christophe Guyeux	Universite de Bourgogne Franche Comté, France
Christophe Soares	University Fernando Pessoa, Portugal
Christos Bouras	University of Patras, Greece

Christos Chrysoulas	London South Bank University, UK
Christos Chrysoulas	Edinburgh Napier University, UK
Ciro Martins	University of Aveiro, Portugal
Claudio Sapateiro	Polytechnic of Setúbal, Portugal
Cosmin Strilechi	Technical University of Cluj-Napoca, Romania
Costin Badica	University of Craiova, Romania
Cristian García Bauza	PLADEMA-UNICEN-CONICET, Argentina
Cristina Caridade	Polytechnic of Coimbra, Portugal
Danish Jamil	Malaysia University of Science and Technology, Malaysia
David Cortés-Polo	University of Extremadura, Spain
David Kelly	University College London, UK
Daria Bylieva	Peter the Great St. Petersburg Polytechnic University, Russia
Dayana Spagnuolo	Vrije Universiteit Amsterdam, Netherlands
Dhouha Jaziri	University of Sousse, Tunisia
Dmitry Frolov	HSE University, Russia
Dulce Mourato	ISTEC - Higher Advanced Technologies Institute Lisbon, Portugal
Edita Butrime	Lithuanian University of Health Sciences, Lithuania
Edna Dias Canedo	University of Brasilia, Brazil
Egils Ginters	Riga Technical University, Latvia
Ekaterina Isaeva	Perm State University, Russia
Eliana Leite	University of Minho, Portugal
Enrique Pelaez	ESPOL University, Ecuador
Eriks Sneiders	Stockholm University, Sweden; Esteban Castellanos ESPE, Ecuador
Fatima Azzahra Amazal	Ibn Zohr University, Morocco
Fernando Bobillo	University of Zaragoza, Spain
Fernando Molina-Granja	National University of Chimborazo, Ecuador
Fernando Moreira	Portucalense University, Portugal
Fernando Ribeiro	Polytechnic Castelo Branco, Portugal
Filipe Caldeira	Polytechnic of Viseu, Portugal
Filippo Neri	University of Naples, Italy
Firat Bestepe	Republic of Turkey Ministry of Development, Turkey
Francesco Bianconi	Università degli Studi di Perugia, Italy
Francisco García-Peñalvo	University of Salamanca, Spain
Francisco Valverde	Universidad Central del Ecuador, Ecuador
Frederico Branco	University of Trás-os-Montes e Alto Douro, Portugal
Galim Vakhitov	Kazan Federal University, Russia

Gayo Diallo	University of Bordeaux, France
Gabriel Pestana	Polytechnic Institute of Setubal, Portugal
Gema Bello-Orgaz	Universidad Politecnica de Madrid, Spain
George Suciu	BEIA Consult International, Romania
Ghani Albaali	Princess Sumaya University for Technology, Jordan
Gian Piero Zarri	University Paris-Sorbonne, France
Giovanni Buonanno	University of Calabria, Italy
Gonçalo Paiva Dias	University of Aveiro, Portugal
Goreti Marreiros	ISEP/GECAD, Portugal
Habiba Drias	University of Science and Technology Houari Boumediene, Algeria
Hafed Zarzour	University of Souk Ahras, Algeria
Haji Gul	City University of Science and Information Technology, Pakistan
Hakima Benali Mellah	Cerist, Algeria
Hamid Alasadi	Basra University, Iraq
Hatem Ben Sta	University of Tunis at El Manar, Tunisia
Hector Fernando Gomez Alvarado	Universidad Tecnica de Ambato, Ecuador
Hector Menendez	King's College London, UK
Hélder Gomes	University of Aveiro, Portugal
Helia Guerra	University of the Azores, Portugal
Henrique da Mota Silveira	University of Campinas (UNICAMP), Brazil
Henrique S. Mamede	University Aberta, Portugal
Henrique Vicente	University of Évora, Portugal
Hicham Gueddah	University Mohammed V in Rabat, Morocco
Hing Kai Chan	University of Nottingham Ningbo China, China
Igor Aguilar Alonso	Universidad Nacional Tecnológica de Lima Sur, Peru
Inês Domingues	University of Coimbra, Portugal
Isabel Lopes	Polytechnic of Bragança, Portugal
Isabel Pedrosa	Coimbra Business School - ISCAC, Portugal
Isaías Martins	University of Leon, Spain
Issam Moghrabi	Gulf University for Science and Technology, Kuwait
Ivan Armuelles Voinov	University of Panama, Panama
Ivan Dunder	University of Zagreb, Croatia
Ivone Amorim	University of Porto, Portugal
Jaime Diaz	University of La Frontera, Chile
Jan Egger	IKIM, Germany
Jan Kubicek	Technical University of Ostrava, Czech Republic
Jeimi Cano	Universidad de los Andes, Colombia

Jesús Gallardo Casero	University of Zaragoza, Spain
Jezreel Mejia	CIMAT, Unidad Zacatecas, Mexico
Jikai Li	The College of New Jersey, USA
Jinzhi Lu	KTH-Royal Institute of Technology, Sweden
Joao Carlos Silva	IPCA, Portugal
João Manuel R. S. Tavares	University of Porto, FEUP, Portugal
João Paulo Pereira	Polytechnic of Bragança, Portugal
João Reis	University of Aveiro, Portugal
João Reis	University of Lisbon, Portugal
João Rodrigues	University of the Algarve, Portugal
João Vidal de Carvalho	Polytechnic of Porto, Portugal
Joaquin Nicolas Ros	University of Murcia, Spain
John W. Castro	University de Atacama, Chile
Jorge Barbosa	Polytechnic of Coimbra, Portugal
Jorge Buele	Technical University of Ambato, Ecuador; Jorge Gomes University of Lisbon, Portugal
Jorge Oliveira e Sá	University of Minho, Portugal
José Braga de Vasconcelos	Universidade Lusófona, Portugal
Jose M. Parente de Oliveira	Aeronautics Institute of Technology, Brazil
José Machado	University of Minho, Portugal
José Paulo Lousado	Polytechnic of Viseu, Portugal
Jose Quiroga	University of Oviedo, Spain
Jose Silvestre Silva	Academia Military, Portugal
Jose Torres	University Fernando Pessoa, Portugal
Juan M. Santos	University of Vigo, Spain
Juan Manuel Carrillo de Gea	University of Murcia, Spain
Juan Pablo Damato	UNCPBA-CONICET, Argentina
Kalinka Kaloyanova	Sofia University, Bulgaria
Kamran Shaukat	The University of Newcastle, Australia
Katerina Zdravkova	University Ss. Cyril and Methodius, North Macedonia
Khawla Tadist	Morocco
Khalid Benali	LORIA - University of Lorraine, France
Khalid Nafil	Mohammed V University in Rabat, Morocco
Korhan Gunel	Adnan Menderes University, Turkey
Krzysztof Wolk	Polish-Japanese Academy of Information Technology, Poland
Kuan Yew Wong	Universiti Teknologi Malaysia (UTM), Malaysia
Kwanghoon Kim	Kyonggi University, South Korea
Laila Cheikhi	Mohammed V University in Rabat, Morocco
Laura Varela-Candamio	Universidade da Coruña, Spain
Laurentiu Boicescu	E.T.T.I. U.P.B., Romania

Lbtissam Abnane	ENSIAS, Morocco
Lia-Anca Hangan	Technical University of Cluj-Napoca, Romania
Ligia Martinez	CECAR, Colombia
Lila Rao-Graham	University of the West Indies, Jamaica
Liliana Ivone Pereira	Polytechnic of Cávado and Ave, Portugal
Łukasz Tomczyk	Pedagogical University of Cracow, Poland
Luis Alvarez Sabucedo	University of Vigo, Spain
Luís Filipe Barbosa	University of Trás-os-Montes e Alto Douro
Luis Mendes Gomes	University of the Azores, Portugal
Luis Pinto Ferreira	Polytechnic of Porto, Portugal
Luis Roseiro	Polytechnic of Coimbra, Portugal
Luis Silva Rodrigues	Polytencic of Porto, Portugal
Mahdieh Zakizadeh	MOP, Iran
Maksim Goman	JKU, Austria
Manal el Bajta	ENSIAS, Morocco
Manuel Antonio Fernández-Villacañas Marín	Technical University of Madrid, Spain
Manuel Ignacio Ayala Chauvin	University Indoamerica, Ecuador
Manuel Silva	Polytechnic of Porto and INESC TEC, Portugal
Manuel Tupia	Pontifical Catholic University of Peru, Peru
Manuel Au-Yong-Oliveira	University of Aveiro, Portugal
Marcelo Mendonça Teixeira	Universidade de Pernambuco, Brazil
Marciele Bernardes	University of Minho, Brazil
Marco Ronchetti	Universita' di Trento, Italy
Mareca María Pilar	Universidad Politécnica de Madrid, Spain
Marek Kvet	Zilinska Univerzita v Ziline, Slovakia
Maria João Ferreira	Universidade Portucalense, Portugal
Maria José Sousa	University of Coimbra, Portugal
María Teresa García-Álvarez	University of A Coruna, Spain
Maria Sokhn	University of Applied Sciences of Western Switzerland, Switzerland
Marijana Despotovic-Zratic	Faculty Organizational Science, Serbia
Marilio Cardoso	Polytechnic of Porto, Portugal
Mário Antunes	Polytechnic of Leiria & CRACS INESC TEC, Portugal
Marisa Maximiano	Polytechnic Institute of Leiria, Portugal
Marisol Garcia-Valls	Polytechnic University of Valencia, Spain
Maristela Holanda	University of Brasilia, Brazil
Marius Vochin	E.T.T.I. U.P.B., Romania
Martin Henkel	Stockholm University, Sweden
Martín López Nores	University of Vigo, Spain
Martin Zelm	INTEROP-VLab, Belgium

Mazyar Zand	MOP, Iran
Mawloud Mosbah	University 20 Août 1955 of Skikda, Algeria
Michal Adamczak	Poznan School of Logistics, Poland
Michal Kvet	University of Zilina, Slovakia
Miguel Garcia	University of Oviedo, Spain
Mircea Georgescu	Al. I. Cuza University of Iasi, Romania
Mirna Muñoz	Centro de Investigación en Matemáticas A.C., Mexico
Mohamed Hosni	ENSIAS, Morocco
Monica Leba	University of Petrosani, Romania
Nadesda Abbas	UBO, Chile
Narasimha Rao Vajjhala	University of New York Tirana, Tirana
Narjes Benameur	Laboratory of Biophysics and Medical Technologies of Tunis, Tunisia
Natalia Grafeeva	Saint Petersburg University, Russia
Natalia Miloslavskaya	National Research Nuclear University MEPhI, Russia
Naveed Ahmed	University of Sharjah, United Arab Emirates
Neeraj Gupta	KIET group of institutions Ghaziabad, India
Nelson Rocha	University of Aveiro, Portugal
Nikola S. Nikolov	University of Limerick, Ireland
Nicolas de Araujo Moreira	Federal University of Ceara, Brazil
Nikolai Prokopyev	Kazan Federal University, Russia
Niranjan S. K.	JSS Science and Technology University, India
Noemi Emanuela Cazzaniga	Politecnico di Milano, Italy
Noureddine Kerzazi	Polytechnique Montréal, Canada
Nuno Melão	Polytechnic of Viseu, Portugal
Nuno Octávio Fernandes	Polytechnic of Castelo Branco, Portugal
Nuno Pombo	University of Beira Interior, Portugal
Olga Kurasova	Vilnius University, Lithuania
Olimpiu Stoicuta	University of Petrosani, Romania
Patricia Quesado	Polytechnic of Cávado and Ave, Portugal
Patricia Zachman	Universidad Nacional del Chaco Austral, Argentina
Paula Serdeira Azevedo	University of Algarve, Portugal
Paula Dias	Polytechnic of Guarda, Portugal
Paulo Alejandro Quezada Sarmiento	University of the Basque Country, Spain
Paulo Maio	Polytechnic of Porto, ISEP, Portugal
Paulvanna Nayaki Marimuthu	Kuwait University, Kuwait
Paweł Karczmarek	The John Paul II Catholic University of Lublin, Poland

Pedro Rangel Henriques	University of Minho, Portugal
Pedro Sobral	University Fernando Pessoa, Portugal
Pedro Sousa	University of Minho, Portugal
Philipp Jordan	University of Hawaii at Manoa, USA
Piotr Kulczycki	Systems Research Institute, Polish Academy of Sciences, Poland
Prabhat Mahanti	University of New Brunswick, Canada
Rabia Azzi	Bordeaux University, France
Radu-Emil Precup	Politehnica University of Timisoara, Romania
Rafael Caldeirinha	Polytechnic of Leiria, Portugal
Raghuraman Rangarajan	Sequoia AT, Portugal
Radhakrishna Bhat	Manipal Institute of Technology, India
Raiani Ali	Hamad Bin Khalifa University, Qatar
Ramadan Elaiees	University of Benghazi, Libya
Ramayah T.	Universiti Sains Malaysia, Malaysia
Ramazy Mahmoudi	University of Monastir, Tunisia
Ramiro Gonçalves	University of Trás-os-Montes e Alto Douro & INESC TEC, Portugal
Ramon Alcarria	Universidad Politécnica de Madrid, Spain
Ramon Fabregat Gesa	University of Girona, Spain
Ramy Rahimi	Chungnam National University, South Korea
Reiko Hishiyama	Waseda University, Japan
Renata Maria Maracho	Federal University of Minas Gerais, Brazil
Renato Toasa	Israel Technological University, Ecuador
Reyes Juárez Ramírez	Universidad Autonoma de Baja California, Mexico
Rocío González-Sánchez	Rey Juan Carlos University, Spain
Rodrigo Franklin Frogeri	University Center of Minas Gerais South, Brazil
Ruben Pereira	ISCTE, Portugal
Rui Alexandre Castanho	WSB University, Poland
Rui S. Moreira	UFP & INESC TEC & LIACC, Portugal
Rustam Burnashev	Kazan Federal University, Russia
Saeed Salah	Al-Quds University, Palestine
Said Achchab	Mohammed V University in Rabat, Morocco
Sajid Anwar	Institute of Management Sciences Peshawar, Pakistan
Sami Habib	Kuwait University, Kuwait
Samuel Sepulveda	University of La Frontera, Chile
Sara Luis Dias	Polytechnic of Cávado and Ave, Portugal
Sandra Costanzo	University of Calabria, Italy
Sandra Patricia Cano Mazuera	University of San Buenaventura Cali, Colombia
Sassi Sassi	FSJEGJ, Tunisia

Seppo Sirkemaa	University of Turku, Finland
Sergio Correia	Polytechnic of Portalegre, Portugal
Shahnawaz Talpur	Mehran University of Engineering & Technology Jamshoro, Pakistan
Shakti Kundu	Manipal University Jaipur, Rajasthan, India
Shashi Kant Gupta	Eudoxia Research University, USA
Silviu Vert	Politehnica University of Timisoara, Romania
Simona Mirela Riurean	University of Petrosani, Romania
Slawomir Zolkiewski	Silesian University of Technology, Poland
Solange Rito Lima	University of Minho, Portugal
Sonia Morgado	ISCPSI, Portugal
Sonia Sobral	Portucalense University, Portugal
Sorin Zoican	Polytechnic University of Bucharest, Romania
Souraya Hamida	Batna 2 University, Algeria
Stalin Figueroa	University of Alcala, Spain
Sümeyya Ilkin	Kocaeli University, Turkey
Syed Asim Ali	University of Karachi, Pakistan
Syed Nasirin	Universiti Malaysia Sabah, Malaysia
Tatiana Antipova	Institute of Certified Specialists, Russia
TatiannaRosal	University of Trás-os-Montes e Alto Douro, Portugal
Tero Kokkonen	JAMK University of Applied Sciences, Finland
The Thanh Van	HCMC University of Food Industry, Vietnam
Thomas Weber	EPFL, Switzerland
Timothy Asiedu	TIM Technology Services Ltd., Ghana
Tom Sander	New College of Humanities, Germany
Tomasz Kisielewicz	Warsaw University of Technology
Tomaž Klobučar	Jozef Stefan Institute, Slovenia
Toshihiko Kato	University of Electro-communications, Japan
Tuomo Sipola	Jamk University of Applied Sciences, Finland
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Valentim Realinho	Polytechnic of Portalegre, Portugal
Valentina Colla	Scuola Superiore Sant'Anna, Italy
Valerio Stallone	ZHAW, Switzerland
Verónica Vasconcelos	Polytechnic of Coimbra, Portugal
Vicenzo Iannino	Scuola Superiore Sant'Anna, Italy
Vitor Gonçalves	Polytechnic of Bragança, Portugal
Victor Alves	University of Minho, Portugal
Victor Georgiev	Kazan Federal University, Russia
Victor Hugo Medina Garcia	Universidad Distrital Francisco José de Caldas, Colombia
Victor Kaptelinin	Umeå University, Sweden



Viktor Medvedev	Vilnius University, Lithuania
Vincenza Carchiolo	University of Catania, Italy
Waqas Bangyal	University of Gujrat, Pakistan
Wolf Zimmermann	Martin Luther University Halle-Wittenberg, Germany
Yadira Quiñonez	Autonomous University of Sinaloa, Mexico
Yair Wiseman	Bar-Ilan University, Israel
Yassine Drias	University of Algiers, Algeria
Yuhua Li	Cardiff University, UK
Yuwei Lin	University of Roehampton, UK
Zbigniew Suraj	University of Rzeszow, Poland
Zorica Bogdanovic	University of Belgrade, Serbia

# Contents

## **1st Workshop on Data Privacy and Protection in Modern Technologies**

GDPR-Compliant Data Breach Detection: Leveraging Semantic Web and Blockchain .....	3
<i>Kainat Ansar, Mansoor Ahmed, Muhammad Irfan Khalid, and Markus Helfert</i>	
Leveraging Blockchain Technologies for Secure and Efficient Patient Data Management in Disaster Scenarios .....	12
<i>Muhammad Irfan Khalid, Mansoor Ahmed, Kainat Ansar, and Markus Helfert</i>	
Oracles in Blockchain Architectures: A Literature Review on Their Implementation in Complex Multi-organizational Processes .....	22
<i>Xavier Gutierrez and José Herrera</i>	

## **1st Workshop on Railway Operations, Modeling and Safety**

Cost Effective Predictive Railway Track Maintenance .....	35
<i>Sri Harikrishnan, Verena Dörner, and Shahrom Sohi</i>	

## **3rd Workshop on Digital Marketing and Communication, Technologies, and Applications**

The Impact of Using Digital Platforms and Sharing Online Experiences on the Reputation of a Company .....	47
<i>Beatriz Pereira, Gabriela Brás, Elvira Vieira, Ana Pinto Borges, Bruno Miguel Vieira, and Manuel Fonseca</i>	
Activating a Brand Through Digital Marketing: The Case of ‘Os Bonitos’ .....	58
<i>Sara Rocha and Alexandra Leandro</i>	
Social Marketing Importance for the Sustainability of Third Sector Organizations .....	68
<i>Susana M. S. R. Fonseca, Filipe A. P. Duarte, Ana Branca Carvalho, Ana Guia, Maria José Madeira, and Geisa Machado</i>	

<b>The Impact of Process Automation on Employee Performance</b> .....	78
<i>Maria João Luz, Manuel José Serra da Fonseca, Jorge Esparteiro Garcia, and José Gabriel Andrade</i>	
<b>Effect of Social Media on Workplace Procrastination Among Employees in Bosnia and Herzegovina</b> .....	88
<i>Suada Pestek, Almir Pestek, and Amra Kozo</i>	
<b>Challenges of Using E-commerce in Bosnia and Herzegovina from the Perspective of Online Store Owners</b> .....	99
<i>Almir Pestek and Nadija Hadzijamakovic</i>	
<b>Analyzing São Paulo's Place Branding Positioning in Promotional Videos (2017–2019)</b> .....	110
<i>José Gabriel Andrade, Adriano Sampaio, Jorge Esparteiro Garcia, Álvaro Cairrão, and Manuel José Serra da Fonseca</i>	
<b>The Influence of TikTok in Portuguese Millennials' Footwear Consumer Behaviour</b> .....	117
<i>Alexandre Duarte and Luís Albuquerque</i>	
 <b>4th Workshop on Open Learning and Inclusive Education Through Information and Communication Technology</b>	
<b>Promoting Inclusion in the Brazilian Educational Scenario: Actions for Teacher Training</b> .....	129
<i>Cibelle A. H. Amato, Cibele C. da S. Spigel, Gerson O. E. Maitana, Andressa G. Saad, Maria Angelica de P. Couto, and Valéria F. Martins</i>	
 <b>1st Workshop on Environmental Data Analytics</b>	
<b>Impact of Preprocessing Using Substitution on the Performance of Selected NER Models - Methodology</b> .....	141
<i>Miroslav Potočár and Michal Kvet</i>	
<b>Correlation <math>n</math>-ptychs of Multidimensional Datasets</b> .....	151
<i>Adam Dudáš</i>	
<b>Performance Analysis of the Data Aggregation in the Oracle Database</b> .....	161
<i>Michal Kvet</i>	
<b>BipartiteJoin: Optimal Similarity Join for Fuzzy Bipartite Matching</b> .....	171
<i>Ondrej Rozínek, Monika Borkovcova, and Jan Mares</i>	

Scalable Similarity Joins for Fast and Accurate Record Deduplication in Big Data .....	181
<i>Ondrej Rozinek, Monika Borkovcova, and Jan Mares</i>	
Impact of Preprocessing Using Substitution on the Performance of Selected NER Models - Results .....	192
<i>Miroslav Potočár</i>	
Oracle APEX as a Tool for Data Analytics .....	203
<i>Ivan Pastierik</i>	
Phishing Webpage Longevity .....	215
<i>Ivan Skula and Marek Kvet</i>	
<b>1st Workshop on AI in Education</b>	
A Conceptual Architecture for Building Intelligent Applications for Cognitive Support in Dementia Care .....	229
<i>Ana Beatriz Silva and Vítor Duarte dos Santos</i>	
<b>1st Workshop on Artificial Intelligence Models and Artifacts for Business Intelligence Applications</b>	
Improving Customer Service Through the Use of Chatbot at Enma Spa Huancayo, Peru .....	241
<i>Elvis Araujo, Diana Javier, and Daniel Gamarra</i>	
NLP in Requirements Processing: A Content Analysis Based Systematic Literature Mapping .....	251
<i>Bell Manrique-Losada, Fernando Moreira, and Eidher Julián Cadavid</i>	
<b>1st Workshop on The Role of the Technologies in the Research of the Migrations</b>	
“From Letters and Phone Calls to WhatsApp and Social Media: The Evolution of Immigration Communication” .....	263
<i>Jessica Ordóñez Cuenca and Analy Poleth Guamán Carrión</i>	
Visual Ethnographic Analysis of the Transit Migration of Venezuelans in Huaquillas, Ecuador .....	267
<i>Pascual Gerardo García-Macías, Marcel Angel Esquivel-Serrano, and Edison Javier Castillo-Pinta</i>	

Evaluation of the Benefit of Artificial Intelligence for the Development of Microeconomics Competencies ..... 273  
*Luís Rojas and Álvaro Méndez*

Ethnography of Tourism in Saraguro: Exploring the Dynamic Legacy of Sumak Kawsay in Local Culture and Heritage ..... 280  
*Edison Javier Castillo-Pinta, Ochoa Jiménez Diego, and Pascual García-Macías*

**12nd Workshop on Special Interest Group on ICT for Auditing and Accounting**

A Guide to Identifying Artificial Intelligence in ERP Systems in Accounting Functions ..... 287  
*Célia Rocha Santos, Graça Azevedo, and Rui Pedro Marques*

Reshaping the Accountant’s Future in the Era of Emerging Technologies ..... 296  
*Ana Ferreira and Isabel Pedrosa*

Factors Influencing Statutory Auditors’ Perception of the Role of Artificial Intelligence in Auditing ..... 306  
*Joana Nogueira, Davide Ribeiro, and Rui Pedro Marques*

Personal Data Protection and Public Disclosure of Data Relating to Taxpayers Debtors to the Portuguese Tax Authority ..... 317  
*Sara Luís Dias*

Beyond Labels and Barriers: Women’s Ongoing Journey in the Auditing Profession ..... 325  
*Silvia Bernardo, Isabel Pedrosa, and Daniela Monteiro*

Promoting Fiscal Incentives for Urban Regeneration: Local Government Digital Presence ..... 335  
*Ana Arromba Dinis*

**2nd Workshop on Data Mining and Machine Learning in Smart Cities**

Deep Learning Approaches for Socially Contextualized Acoustic Event Detection in Social Media Posts ..... 347  
*Vahid Hajjhashemi, Abdorreza Alavi Gharahbagh, Marta Campos Ferreira, José J. M. Machado, and João Manuel R. S. Tavares*

Abnormal Action Recognition in Social Media Clips Using Deep Learning  
to Analyze Behavioral Change ..... 359  
*Abdorreza Alavi Gharahbagh, Vahid Hajhashemi,  
Marta Campos Ferreira, José J. M. Machado,  
and João Manuel R. S. Tavares*

**2nd Workshop on Enabling Software Engineering Practices Via Last  
Development Trends**

Exploring Software Quality Through Data-Driven Approaches  
and Knowledge Graphs ..... 373  
*Raheela Chand, Saif Ur Rehman Khan, Shahid Hussain, Wen-Li Wang,  
Mei-Huei Tang, and Naseem Ibrahim*

**Author Index** ..... 383



# Impact of Preprocessing Using Substitution on the Performance of Selected NER Models - Results

Miroslav Potočár<sup>(✉)</sup>

University of Žilina, Žilina, Slovakia  
Miroslav.Potocar@fri.uniza.sk

**Abstract.** This paper investigates the impact of preprocessing using substitution in word sequences on selected models of named entity recognition. The study is focused on evaluating the results of the performed experiments. It briefly describes the concept of substitution using pseudo words and the methodology used in performing the experiments. Based on the outputs of the experiments, it discusses in detail the implications of substitution on the selected models and provides possible explanations for the results. In the end, conclusions and recommendations for the use of substitution as a preprocessing technique are made based on the observed results.

**Keywords:** named entity recognition · preprocessing · substitution · pseudo words

## 1 Introduction

Nowadays, the number of data increases considerably over time [2]. Much of this data is in the form of unstructured text, so a lot of attention is currently being paid to natural language processing tasks. Named Entity Recognition (NER) task is focused on identifying, locating and classifying important objects in text data [5, 8]. NER is a fundamental task that is addressed in other tasks such as information extraction, question answering and knowledge ontology construction. In these tasks, NER is indispensable [3].

Data preprocessing is an important aspect in solving many natural language processing tasks. Despite its importance, it has received little attention in the literature. In this paper, we investigate a specific preprocessing strategy - substitution - and its impact on the performance of selected NER models. Substitution is the strategic replacement of words in sequences by pseudo words that encode a certain feature of the replaced word. This type of substitution may reduce overfitting and improve the model's generalization abilities for some models under certain conditions.

This paper is dedicated to revealing the results of a large set of experiments designed to evaluate the impact of preprocessing using substitution on selected NER models. Specifically, we investigate the effects on the hidden Markov model

(HMM), conditional random fields (CRF), gated recurrent unit (GRU), bidirectional long short-term memory network (BiLSTM) and our Naïve model. With respect to the experiment results, we aim to provide a comprehensive understanding of how substitution affects the prediction capabilities of the models.

## 2 Concept of Pseudo Word Substitution

The idea to investigate the impact of word substitution in the sequence by pseudo words arose while studying the work of Bikel et al. [1] where pseudo words appeared as one of the features entering the process of solving the NER task. We took this concept of pseudo words and focused on researching how the use of pseudo words affects NER models whose input is only a sequence of words. The idea is to replace unknown or rarely occurring words with a pseudo word that encodes one of the features. We have taken the word features from the original work, with their order, examples, and the intuition behind each feature, and extended them to include a representative pseudo word and a condition that must be fulfilled in order to replace the word with the pseudo word. We substitute words that:

- consist exclusively of two numbers,
- consist exclusively of four numbers,
- containing numbers and letters,
- contain numbers and dashes,
- contain numbers and slashes,
- contain numbers and commas,
- contain numbers and a period,
- represent other numbers,
- consist entirely of capital letters,
- have a capital initial letter and a period,
- are the first word in a sentence,
- have a capital initial letter,
- consist of lower case letters only,
- contains any alphanumeric and non-alphanumeric characters.

A detailed list of pseudo words along with rules, examples and intuition can be found in [6].

An example of the functioning and impact of substitution is best illustrated with words containing numbers. We often encounter this type of words in texts. If we wanted to include all numbers in the vocabulary, this would not be possible due to their infinite nature. Words such as ‘22/11/2023’ and ‘23/11/2023’, would appear as unique words within the vocabulary, but we can infer from their structure that they are words representing a date. For NER models to be able to identify that a given word is a date, they would need to encounter the particular word in multiple possible contexts. However, in this way they would only learn to recognize one particular date. Clearly, this method of learning would lead to overfitting and an inability to generalize in NER models. If we replace



each word that has the shape of a date with a pseudo word [CDS], the model will have more opportunities to learn the context in which dates occur, and as a result the model will be capable to better handle new, unique dates as well. This improves the model’s ability to generalize and prevents overfitting of the model. Similar logic can be applied to words containing a capital letter. The set of possible company names is theoretically unlimited. However, most names share a common feature, which is the first initial letter. Replacing this name with the pseudo word [IC] allows the model to better learn the context in which the company names may occur.

### 3 Methodology

In the following section we briefly describe the methodology used. A detailed explanation of our experimental procedure can be found in the paper [6].

#### 3.1 Data

As test data, we used the dataset *CoNLLpp* [9], which is a modification of the original dataset *CoNLL2003* [7]. The dataset uses the IOB2 labeling scheme. It distinguishes four types of entities, persons (PER), locations (LOC), organizations (ORG) and miscellaneous (MISC). There are three sets already prepared in the dataset namely training, validation and test sets. A summary of the data in each part of the dataset can be seen in Table 1. For each part, we have listed the number of sentences, the number of words, the number of unique words. For each IOB category, we have listed the number of words associated with this category and also the number of unique words for this category.

#### 3.2 Models Implementations

We have used *Python* in our research, so the model implementations used are influenced by this.

In the case of the Naïve model, we self-implemented a simple class that stored information about words and the named entity tag that occurred most frequently with a given word in the training set. It assigns the most frequent tag in train dataset to unknown words.

For the HMM, we have used the *HiddenMarkovModelTagger* implementation available in *NLTK* library. This implementation can be used in sequence prediction, and is also able to handle unseen tokens.

As a CRF model, we used the *CRF* implementation available in the *sklearn-crfsuite* library, which is a library focused specifically on the CRF model. Unfortunately, this library is no longer active. Because of this, we have encountered several issues when we were using it, related to incompatibility with newer versions of the *numpy* library. However, due to the library being open source, it was possible to fix these issues and make use of the implementation.

To implement the GRU and BiLSTM models, we used the tools provided by the *Keras* and *TensorFlow* libraries.

**Table 1.** Datasets summary

	Train	Validation	Test
Sentences	14041	3250	3453
Words	203621	51362	46435
Unique words	23623	9966	9488
B-LOC words	7140	1837	1646
B-LOC unique words	1223	511	476
B-MISC words	3438	922	723
B-MISC unique words	707	304	279
B-ORG words	6321	1341	1715
B-ORG unique words	1767	547	684
B-PER words	6600	1842	1618
B-PER unique words	2275	919	858
I-LOC words	1157	257	259
I-LOC unique words	294	90	105
I-MISC words	1155	346	254
I-MISC unique words	334	148	115
I-ORG words	3704	751	882
I-ORG unique words	1124	334	399
I-PER words	4528	1307	1161
I-PER unique words	2398	952	797
O words	169578	42759	38177
O unique words	15704	6806	6401

### 3.3 Replacing Words with Pseudo Words

The process of word substitution with pseudo words consists of several steps. At the beginning, a dictionary of known words is created based on the training data. From this dictionary, depending on the scenario, words containing numbers, composed entirely of non-alphanumeric characters, or words whose number of occurrences within the training dataset satisfy a certain threshold are removed or kept. In the next step, the individual words in each of the datasets are sequentially reviewed and it is determined whether they should be replaced by a pseudo word. First, the occurrence of the word in the dictionary of known words is checked. If it is found in this dictionary, the original form is retained and it is discarded from further processing. If it is not found in the dictionary, it means that this word will be replaced by a pseudo word. For words not found in the dictionary of known words, the conditions are successively applied in the exact order. Depending on which condition the word satisfies, it is replaced by the corresponding pseudo word. The datasets modified in this way are used to

train, validate and also test the models. The detailed word substitution process is described in the paper [6].

### 3.4 Test Scenarios

As described in detail in the paper [6], we performed several experimental scenarios with each model. For each model, we determined its raw performance, i.e., the overall F1 score that the model achieved when no preprocessing was being applied. We then moved on to the actual scenarios where preprocessing was used.

We tested two types of preprocessing scenarios. In the first type of scenario, we did not remove words consisting entirely of punctuation or words containing numerical values from the set of all words that were later used to build the dictionary of known words. We only removed those words from the set whose number of occurrences was below a given threshold. Specifically, we focused on scenarios where we removed words that had a frequency of occurrences less than 1, 2, 3, 4, and 5. Each frequency threshold represented one test scenario. In the second type of scenario, we removed words containing numbers or composed entirely of non-alphanumeric characters from the set of all words used to build the dictionary of known words. As in the first type of scenario, we also removed words whose number of occurrences was less than 1, 2, 3, 4, and 5 from the rest of the words. From the remaining words, a dictionary of known words was created and used during the substitution process.

### 3.5 Evaluation

To evaluate the performance of the models, we used the F1 score metric, which is widely used in NER task evaluation. F1 score is a combination of precision and recall metrics. Thus, it expresses the balance of both metrics and indicates how well the model is able to correctly identify entities and at the same time what is the ability of the model to detect all real entities.

We used the *segeval* [4] framework available through the *evaluate* library, which is designed to evaluate labeled sequences. This framework provides values for precision, accuracy, recall, and F1 score for the entire dataset and also provides the same metrics with respect to specific categories of named entities. The *segeval* provides two evaluation modes, **default** and **strict**. The default mode simulates *conllev* and **strict** evaluates the inputs based on a specific schema. Our dataset used the IOB2 schema, so we used the **strict** mode where we defined the use of the IOB2 schema.

For each test scenario, we performed 5 runs, which means that we recreated and retrained the model 5 times and evaluated its performance on each dataset. The final value of each metric is calculated as the average of the measured values in each run.

## 4 Results

The results of the experiments are shown in Table 2. One row represents the averaged results for one tested scenario. Each row consists of the following values:

- **Model** - This column specifies the tested model. In our case, we performed the experiments with the Naïve, HMM, CRF, GRU and BiLSTM models.
- **S** - This column provides information on if substitution was (T) or was not (F) performed over the individual datasets.
- **N&P** - This column contains information about whether words that contained a number and words that consisted entirely of punctuation were removed (T) or kept (F) in a given scenario.
- **FT** - This column informs about the threshold value of the frequency of occurrence. Words that had a frequency of occurrence equal or less than this value were excluded from the dictionary of known words.
- **P, L, O, M** - Provides the F1 score for the specified named entity type (P = PER, L = LOC, O = ORG, M = MISC).
- $\Delta P$ ,  $\Delta L$ ,  $\Delta O$ ,  $\Delta M$  - Indicates how the F1 score for a given named entity type has changed from the value without substitution (P = PER, L = LOC, O = ORG, M = MISC).
- **OA** - Indicates the overall value of the F1 score.
- $\Delta$  - Indicates how the overall F1 score changed from the value without substitution.

The grey rows indicate the scenarios where the best result was achieved for a particular model with respect to the overall F1 score.

Considering the increasing trend of the overall F1 score with increasing threshold frequency of occurrence for the BiLSTM model, we performed an additional set of experiments. In these experiments, we observed how the model performed when substitution was used, words containing numbers and words consisting entirely of punctuation were removed, and the threshold varied from 3 to 6. We combined the measured values with the original values for the BiLSTM model and calculated the average values in this case as well. All values were calculated based on 5 runs. The exceptions were the scenarios with a threshold equal to 3 and 4, where the value was calculated based on 10 runs (5 original runs and 5 additional runs). We called the model *BiLSTM-6* and the resulting values are shown in Table 3.

Table 4 contains a subset of the rows from Table 2, giving for each model the row where the best overall F1 scores were obtained. It also includes a row from Table 3, for the BiLSTM model that was tested on an additional set of experiments. We have created an additional Table 5 to this Table 4, which for each model with the best F1 score on the test set, shows how that model performed on the training data.

The Table 4 shows that the use of substitution led to an improvement in F1 scores on the test data for each of the tested models. The exception may be the CRF model, where the difference between the overall F1 score for the with and without substitution scenario is very small. Considering the Table 2, we can see

**Table 2.** Experiment results - test data predictions

Model	Processing			F1 score per class(%)								F1 score(%)	
	S	N&P	FT	P	$\Delta P$	L	$\Delta L$	O	$\Delta O$	M	$\Delta M$	OA	$\Delta$
Naïve	F	F	0	21.68	0.0	75.22	0.0	54.05	0.0	66.12	0.0	54.62	0.0
Naïve	T	F	1	45.96	24.27	75.25	0.03	51.54	-2.51	64.16	-1.96	59.8	5.18
Naïve	T	F	2	44.49	22.81	74.66	-0.56	46.14	-7.92	62.75	-3.36	57.62	3.01
Naïve	T	F	3	40.48	18.79	73.28	-1.94	44.52	-9.53	60.21	-5.91	55.4	0.79
Naïve	T	F	4	37.12	15.43	71.7	-3.52	39.3	-14.75	56.7	-9.42	52.24	-2.37
Naïve	T	T	0	21.68	0.0	75.22	0.0	53.9	-0.16	66.12	0.0	54.57	-0.04
Naïve	T	T	1	45.96	24.27	75.25	0.03	51.38	-2.68	64.16	-1.96	59.75	5.14
Naïve	T	T	2	44.49	22.81	74.66	-0.56	46.04	-8.02	62.75	-3.36	57.6	2.98
Naïve	T	T	3	40.48	18.79	73.28	-1.94	44.42	-9.64	60.21	-5.91	55.38	0.76
Naïve	T	T	4	37.12	15.43	71.7	-3.52	39.19	-14.87	56.7	-9.42	52.22	-2.4
HMM	F	F	0	67.72	0.0	66.81	0.0	60.98	0.0	55.17	0.0	63.71	0.0
HMM	T	F	1	74.8	7.08	76.75	9.94	65.31	4.33	69.12	13.95	72.06	8.34
HMM	T	F	2	72.76	5.04	80.96	14.14	60.74	-0.24	66.15	10.98	70.45	6.73
HMM	T	F	3	71.23	3.51	79.39	12.58	59.1	-1.88	63.84	8.67	68.75	5.03
HMM	T	F	4	71.54	3.82	78.05	11.23	55.75	-5.23	61.59	6.42	67.04	3.33
HMM	T	T	0	71.38	3.66	80.4	13.59	64.16	3.18	58.91	3.74	70.32	6.61
HMM	T	T	1	75.87	8.15	76.92	10.1	64.56	3.58	67.48	12.32	72.02	8.3
HMM	T	T	2	64.74	-2.98	81.01	14.2	60.3	-0.68	66.87	11.71	68.19	4.47
HMM	T	T	3	63.42	-4.3	79.46	12.65	58.27	-2.71	64.4	9.24	66.41	2.69
HMM	T	T	4	65.35	-2.37	77.9	11.09	51.05	-9.93	62.11	6.94	63.27	-0.44
CRF	F	F	0	83.32	0.0	84.05	0.0	72.36	0.0	75.87	0.0	79.48	0.0
CRF	T	F	1	82.2	-1.12	82.61	-1.44	69.79	-2.57	72.26	-3.61	77.49	-1.99
CRF	T	F	2	83.2	-0.12	82.36	-1.7	67.96	-4.4	73.89	-1.98	77.37	-2.11
CRF	T	F	3	82.38	-0.95	82.51	-1.54	67.0	-5.36	72.93	-2.94	76.79	-2.69
CRF	T	F	4	82.04	-1.28	83.24	-0.81	67.66	-4.7	71.56	-4.31	76.91	-2.57
CRF	T	T	0	35.93	-47.39	84.56	0.51	69.61	-2.76	74.27	-1.6	66.41	-13.07
CRF	T	T	1	79.84	-3.48	85.0	0.95	72.65	0.28	71.75	-4.12	78.41	-1.08
CRF	T	T	2	79.82	-3.5	86.15	2.1	73.43	1.07	72.71	-3.15	79.07	-0.42
CRF	T	T	3	80.21	-3.11	86.29	2.23	74.74	2.38	72.77	-3.1	79.58	0.09
CRF	T	T	4	79.7	-3.62	85.96	1.91	75.15	2.79	71.05	-4.81	79.23	-0.25
GRU	F	F	0	66.51	0.0	75.52	0.0	59.07	0.0	62.88	0.0	67.03	0.0
GRU	T	F	1	73.87	7.36	79.34	3.82	62.07	2.99	64.73	1.85	70.94	3.91
GRU	T	F	2	74.94	8.43	78.86	3.34	59.14	0.06	62.59	-0.29	69.92	2.89
GRU	T	F	3	74.03	7.52	77.92	2.4	57.75	-1.32	61.32	-1.56	68.97	1.94
GRU	T	F	4	72.53	6.03	76.52	1.0	55.32	-3.75	58.89	-3.98	67.26	0.22
GRU	T	T	0	57.28	-9.23	77.83	2.3	62.44	3.37	62.7	-0.17	64.82	-2.21
GRU	T	T	1	74.61	8.1	78.55	3.02	63.35	4.28	65.25	2.37	71.39	4.36
GRU	T	T	2	74.34	7.83	78.85	3.33	60.82	1.75	63.48	0.6	70.42	3.39
GRU	T	T	3	73.4	6.89	77.18	1.66	60.58	1.51	62.98	0.1	69.65	2.62
GRU	T	T	4	71.45	4.95	76.18	0.66	56.3	-2.77	60.01	-2.87	67.14	0.11
BiLSTM	F	F	0	66.44	0.0	75.22	0.0	65.12	0.0	67.29	0.0	69.07	0.0
BiLSTM	T	F	1	78.59	12.15	82.46	7.23	70.85	5.73	65.56	-1.74	75.84	6.78
BiLSTM	T	F	2	79.29	12.85	82.18	6.96	67.35	2.23	65.08	-2.21	74.79	5.72
BiLSTM	T	F	3	79.62	13.17	82.12	6.9	68.6	3.49	65.78	-1.51	75.3	6.23
BiLSTM	T	F	4	79.12	12.68	81.61	6.38	69.12	4.01	63.5	-3.8	74.8	5.74
BiLSTM	T	T	0	72.52	6.08	81.23	6.01	66.13	1.02	68.42	1.13	72.95	3.88
BiLSTM	T	T	1	76.46	10.02	83.78	8.56	68.41	3.29	66.58	-0.71	75.16	6.09
BiLSTM	T	T	2	77.14	10.7	83.91	8.69	70.05	4.94	67.64	0.35	75.98	6.91
BiLSTM	T	T	3	78.02	11.57	83.6	8.38	71.16	6.04	65.81	-1.48	76.2	7.13
BiLSTM	T	T	4	78.02	11.58	84.02	8.79	71.83	6.72	65.52	-1.77	76.42	7.35

in the CRF case that the application of substitution led to a deterioration in performance on the test set in all scenarios except one.

Referring to the Table 4, we can notice the major change compared to the scenario without substitution in the case of the HMM model. There is an 8.34% increase in the F1 score. Significant improvements also occurred for GRU, BiLSTM and a non-negligible improvement also occurred for our Naïve model.

Looking at the Table 4, we see that for BiLSTM, the overall F1 score increases as the threshold increases. An additional set of experiments presented in the Table 3 indicate that this trend is random and ends at a threshold of 3.

**Table 3.** Additional BiLSTM experiments - test data predictions

Model	Processing			F1 score per class(%)								F1 score(%)	
	S	N&P	FT	P	$\Delta P$	L	$\Delta L$	O	$\Delta O$	M	$\Delta M$	OA	$\Delta$
BiLSTM-6	F	F	0	66.44	0.0	75.22	0.0	65.12	0.0	67.29	0.0	69.07	0.0
BiLSTM-6	T	T	0	72.52	6.08	81.23	6.01	66.13	1.02	68.42	1.13	72.95	3.88
BiLSTM-6	T	T	1	76.46	10.02	83.78	8.56	68.41	3.29	66.58	-0.71	75.16	6.09
BiLSTM-6	T	T	2	77.14	10.7	83.91	8.69	70.05	4.94	67.64	0.35	75.98	6.91
BiLSTM-6	T	T	3	77.99	11.55	83.95	8.73	71.29	6.18	66.22	-1.08	76.36	7.29
BiLSTM-6	T	T	4	77.91	11.46	84.22	9.0	71.6	6.49	65.38	-1.91	76.34	7.28
BiLSTM-6	T	T	5	77.42	10.98	83.81	8.59	71.19	6.08	65.18	-2.11	75.93	6.87
BiLSTM-6	T	T	6	75.5	9.06	83.64	8.42	70.21	5.09	62.3	-5.0	74.51	5.44

**Table 4.** Best overall F1 model - test data predictions

Model	Processing			F1 score per class(%)								F1 score(%)	
	S	N&P	FT	P	$\Delta P$	L	$\Delta L$	O	$\Delta O$	M	$\Delta M$	OA	$\Delta$
Naïve	T	F	1	45.96	24.27	75.25	0.03	51.54	-2.51	64.16	-1.96	59.8	5.18
HMM	T	F	1	74.8	7.08	76.75	9.94	65.31	4.33	69.12	13.95	72.06	8.34
CRF	T	T	3	80.21	-3.11	86.29	2.23	74.74	2.38	72.77	-3.1	79.58	0.09
GRU	T	T	1	74.61	8.1	78.55	3.02	63.35	4.28	65.25	2.37	71.39	4.36
BiLSTM	T	T	4	78.02	11.58	84.02	8.79	71.83	6.72	65.52	-1.77	76.42	7.35
BiLSTM-6	T	T	3	77.99	11.55	83.95	8.73	71.29	6.18	66.22	-1.08	76.36	7.29

When we look at the Table 4 and Table 5 we see that the overall F1 score of the best models on the training set is significantly higher than that on the test set. This indicates that the models are overfitted. The decrease in the overall F1 score on the training data indicates that the use of substitution helps to reduce overfitting and improves the ability of the model to generalize.

**Table 5.** Models with the best overall F1 score on test data - train data predictions

Model	Processing			F1 score per class(%)								F1 score(%)	
	S	N&P	FT	P	$\Delta P$	L	$\Delta L$	O	$\Delta O$	M	$\Delta M$	OA	$\Delta$
Naïve	T	F	1	69.56	-5.69	85.22	-3.79	74.33	-5.45	78.2	-4.4	76.93	-4.8
HMM	T	F	1	91.84	-1.53	90.46	-4.17	85.51	-3.3	87.26	-0.4	89.06	-2.63
CRF	T	T	3	93.57	-6.12	95.45	-3.97	92.17	-6.97	92.82	-5.86	93.66	-5.65
GRU	T	T	1	95.33	-2.56	92.59	-2.27	87.64	-4.39	87.94	-3.84	91.37	-3.14
BiLSTM	T	T	4	93.21	-5.26	95.48	-2.59	90.26	-6.54	88.98	-6.67	92.5	-4.99
BiLSTM-6	T	T	3	94.5	-3.97	96.21	-1.86	92.25	-4.55	90.17	-5.48	93.8	-3.69

Considering the results in the Table 4, we see that most models performed best when words containing number or consisting only of punctuation were replaced by pseudo words (N&P is T). The exceptions are the Naïve and HMM models, where the model performed best when such words were kept. However, when we look at the Table 2, we see that the difference between the best model with omitted and retained words is very small for the Naïve and HMM models.

From the Table 4, we can infer that in most cases, the substitution of words that occurred only once in the training data (FT = 1) has the greatest benefit. We can notice this for models that, during tag prediction for a processed word, do not have access to the words that follow after it (Naïve, HMM, GRU). In contrast, words that can peek at least one word into the future (CRF, BiLSTM) perform best when more frequent words are replaced by pseudo words (FT > 1).

## 5 Discussion and Conclusions

Considering the results of our experiments, we conclude that word substitution using pseudo words has a beneficial impact on the performance of most of the tested models. The models for which we applied the substitution before the training phase performed better on the test data in terms of F1 scores.

The most significant difference appeared in the HMM model, where an 8.34% increase in F1 score was observed compared to the no-substitution scenario. The least increase was observed in the CRF model, where the F1 score rose by 0.09%. We suggest that substitution has the greatest impact on models that do not use sophisticated features to predict tags in a sequence, relying only on the current word being processed, or using knowledge about the position of that word in the sentence. For these models, pseudo words supply a significant amount of information that they use in their predictions. We consider this to be the reason why the CRF model did not achieve significant improvement. CRF also made its predictions from the beginning based on features, which were, for example, information about whether a word consists entirely of capital letters, whether it has a capitalized first letter, whether it is a number, whether it is at the beginning or at the end of a sentence. Much of this information was encoded just through the pseudowords we used. For the CRF model, the pseudo words did not provide

any additional information. Indeed, by replacing the word with a pseudo word, the model lost information such as the last 2 and 3 characters of the word. This assumption is to some extent confirmed by the values in Table 2, where for the CRF model we see for almost every scenario a deterioration compared to the case without substitution. The improvement in one of the cases may be due to coincidence rather than causality.

When we compare the F1 scores on the training and test data, we see that there is overfitting in the models. After applying the substitution, the difference in F1 scores is reduced. Thus, substitution appears to be a good way to reduce model overfitting for some models and also as a way to improve the ability of the model to generalize.

Considering the dataset used, the types of named entities recognized within it, and the pseudowords along with the intuition behind them, we suggest that substitution may improve the prediction ability for similar entities in certain models. Based on the intuition behind the pseudo words used, we also predict a possible improvement if dates, monetary amounts, product names, etc. will be recognized in the text. However, the currently used list of pseudo words emphasizes primarily on features such as letter case and the presence of numbers in a specific form. These may not provide useful information in domains where the recognized named entities do not have these features. We assume that in such domains, substitution would lead to a deterioration of the overall performance for some models.

In future research, we propose to investigate the impact of pseudo word substitution on datasets in different languages and domains. We also propose to investigate the impact of pseudo word substitution in the case of models that will not suffer from overfitting. Some attention should also be given to the design of new pseudo words that would be able to more finely discriminate between cases and also encode additional information.

**Acknowledgment.** It was supported by the Erasmus+ project: Project number: 2022-1-SK01-KA220-HED-000089149, Project title: Including EVERYone in GREEN Data Analysis (EVERGREEN) funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Slovak Academic Association for International Cooperation (SAAIC). Neither the European Union nor SAAIC can be held responsible for them.



## References

1. Bikel, D.M., Schwartz, R., Weischedel, R.M.: An algorithm that learns what's in a name. *Mach. Learn.* **34**, 211–231 (1999)
2. Kvet, M.: Relational data index consolidation. In: 2021 28th Conference of Open Innovations Association (FRUCT), pp. 215–221. IEEE (2021)



3. Li, X., Wang, T., Pang, Y., Han, J., Shi, J.: Review of research on named entity recognition. In: Sun, X., Zhang, X., Xia, Z., Bertino, E. (eds.) ICAIS 2022. CCIS, vol. 1587, pp. 256–267. Springer, Cham (2022). [https://doi.org/10.1007/978-3-031-06761-7\\_21](https://doi.org/10.1007/978-3-031-06761-7_21)
4. Nakayama, H.: seqeval: a python framework for sequence labeling evaluation (2018). <https://github.com/chakki-works/seqeval>
5. Pakhale, K.: Comprehensive overview of named entity recognition: models, domain-specific applications and challenges. arXiv preprint [arXiv:2309.14084](https://arxiv.org/abs/2309.14084) (2023)
6. Potočár, M., Kvet, M.: Impact of preprocessing using substitution on the performance of selected ner models - methodology (2023, manuscript submitted for publication)
7. Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: language-independent named entity recognition. arXiv preprint [cs/0306050](https://arxiv.org/abs/cs/0306050) (2003)
8. Wang, Y., Zhao, W., Wan, Y., Deng, Z., Yu, P.S.: Named entity recognition via machine reading comprehension: a multi-task learning approach. arXiv preprint [arXiv:2309.11027](https://arxiv.org/abs/2309.11027) (2023)
9. Wang, Z., Shang, J., Liu, L., Lu, L., Liu, J., Han, J.: Crossweigh: training named entity tagger from imperfect annotations. arXiv preprint [arXiv:1909.01441](https://arxiv.org/abs/1909.01441) (2019)