LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

**The 4th Workshop on Open-Source Arabic Corpora and Processing Tools with a Shared Task on Offensive Language Detection**

# PROCEEDINGS

Editors:

Hend Al-Khalifa, Walid Magdy, Kareem Darwish, Tamer Elsayed and Hamdy Mubarak

# Proceedings of the LREC 2020 4th Workshop on Open-Source Arabic Corpora and Processing Tools

# with a Shared Task on Offensive Language Detection (OSACT 2020)

Edited by: Hend Al-Khalifa, Walid Magdy, Kareem Darwish, Tamer Elsayed and Hamdy Mubarak

**ISBN: 979-10-95546-51-1**
**EAN: 9791095546511**

**For more information:**
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

# Preface

Given the success of the first, second, and third workshops on Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) in LREC 2014, LREC 2016 and LREC 2018, the fourth workshop comes to encourage researchers and practitioners of Arabic language technologies, including computational linguistics (CL), natural language processing (NLP), and information retrieval (IR) to share and discuss their research efforts, corpora, and tools. The workshop gives special attention to Human Language Technologies based on AI/Machine Learning, which is one of LREC 2020 hot topics. In addition to the general topics of CL, NLP and IR, the workshop gives special emphasis to Offensive Language Detection shared task.

OSACT4 had an acceptance rate of 50%, where we received 12 regular papers from which 6 papers were accepted, in addition to 11 shared task papers. We believe that the accepted papers are of high quality and present a mixture of interesting topics.

This year, we introduced the Shared Task on Offensive Language Detection. The shared task attempts to detect such speech in the realm of Arabic social media in two Subtasks. In subtask A, SemEval 2020 Arabic offensive language dataset, which contains 10,000 tweets that were manually annotated for offensiveness was used. Offensive tweets contain explicit or implicit insults or attacks against other people, or inappropriate language. Subtask B targets the identification of Hate Speech. A hate speech tweet contains insults or threats targeting a group based on their nationality, ethnicity, gender, political or sport affiliation, religious belief, or other common characteristics. Subtasks A and B share the same data splits. Subtask B is more challenging than Subtask A as only 5% of the tweets are labeled as hate speech, while 20% of the tweets are labeled as offensive. The shared task attracted many teams from different countries in the Middle East, Europe and US. In all, 40 and 33 teams signed up to participate in Subtasks A and B; among them, 14 and 13 teams submitted their system outputs in the two subtasks respectively.

We would like to thank everyone who in one way or another helped in making this workshop a success. Our special thanks go to the members of the program committee, who did an excellent job in reviewing the submitted papers, and to the LREC organizers. Last but not least, we would like to thank our authors and the workshop participants.

Hend Al-Khalifa, Walid Magdy, Kareem Darwish, Tamer Elsayed, and Hamdy Mubarak

OSACT4 Organizing Committee

**Organizers:**

Hend Al-Khalifa, King Saud University, KSA
Walid Magdy, University of Edinburgh, UK
Kareem Darwish, Qatar Computing Research Institute, Qatar
Tamer Elsayed, Qatar University, Qatar
Hamdy Mubarak, Qatar Computing Research Institute, Qatar

**Program Committee:**

Nizar Habash, New York University Abu Dhabi, UAE
Wajdi Zaghouani, Hamad Bin Khalifa University, Qatar
Wassim El-Hajj, American University of Beirut, Lebanon
Ayah Zirikly, George Washington University, USA
Irina Temnikova, Sofia, Sofia City Province, Bulgaria
Shady Elbassuoni, American University of Beirut, Lebanon
Nora Al-Twairesh, King Saud University, KSA
Abeer Aldayel, University of Edinburgh, UK
Khaled Shaalan, The British University in Dubai, UAE
Almoataz B. Elsaid, Cairo University, Egypt
Ahmed Mourad, RMIT University, Australia
Hassan Sawaf, Amazon, USA
Fethi Bougares, Université du Maine, Avenue Laënnec, France
Nada Ghneim, Higher Institute for Applied Science and Technology, Syria
Maha Althobaiti, Taif University, KSA
Nasser Zalmout, New York University Abu Dhabi, UAE
Mohammad Salameh, University of Alberta, Canada
Alexis Nasr, Université Aix Marseille, France
AbdelRahim Elmadany, The University of British Columbia, Canada
Mohamed Abdelmageed, The University of British Columbia, Canada
Ahmed Ali, Qatar Computing Research Institute, Qatar
Haithem Afli, Cork Institute of Technology, Ireland
Preslav Nakov, Qatar Computing Research Institute, Qatar
Fahim Dalvi, Qatar Computing Research Institute, Qatar
Salam Khalifa, NYU-AD, UAE
Hassan Sajjad, Qatar Computing Research Institute, Qatar
Maha Alamri, Bangor University, UK
Sarah Kohail, University of Hamburg, Germany
Azzeddine Mazroui, Université Mohammed Premier, Morocco
Bassam Haddad, University of Petra, Jordan
Younes Samih, Qatar Computing Research Institute, Qatar
Khaled Shaban, Qatar University, Qatar
Reem Suwaileh, Qatar University, Qatar
Mucahid Kutlu, TOBB University, Turkey
Maram Hasanain , Qatar University, Qatar
Raghad Alshalaan, Imam Abdulrahman Bin Faisal University, KSA
Shahad Alshalaan, Imam Abdulrahman Bin Faisal University, KSA
Maha Alrabiah, Al Imam Mohammad Ibn Saud Islamic, KSA
Ibrahim Abu Farha, University of Edinburgh, UK

# Table of Contents

# Workshop Program

*An Arabic Tweets Sentiment Analysis Dataset (ATSAD) using Distant Supervision and Self Training*
Kathrein Abu Kwaik, Stergios Chatzikyriakidis, Simon Dobnik, Motaz Saad and Richard Johansson

*AraBERT: Transformer-based Model for Arabic Language Understanding*
wissam antoun, Fady Baly and Hazem Hajj

*AraNet: A Deep Learning Toolkit for Arabic Social Media*
Muhammad Abdul-Mageed, Chiyu Zhang, Azadeh Hashemi and El Moatez Billah Nagoudi

*Building a Corpus of Qatari Arabic Expressions*
Sara Al-Mulla and Wajdi Zaghouani

*From Arabic Sentiment Analysis to Sarcasm Detection: The ArSarcasm Dataset*
Ibrahim Abu Farha and Walid Magdy

*Understanding and Detecting Dangerous Speech in Social Media*
Ali Alshehri, El Moatez Billah Nagoudi and Muhammad Abdul-Mageed

*Overview of OSACT4 Arabic Offensive Language Detection Shared Task*
Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed and Hend Al-Khalifa

*OSACT4 Shared Task on Offensive Language Detection: Intensive Preprocessing-Based Approach*
Fatemah Husain

*ALT Submission for OSACT Shared Task on Offensive Language Detection*
Sabit Hassan, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, Ammar Rashed and Shammur Absar Chowdhury

*ASU_OPTO at OSACT4 - Offensive Language Detection for Arabic text*
Amr Keleg, Samhaa R. El-Beltagy and Mahmoud Khalil

*OSACT4 Shared Tasks: Ensembled Stacked Classification for Offensive and Hate Speech in Arabic Tweets*
Hafiz Hassaan Saeed, Toon Calders and Faisal Kamiran

*Arabic Offensive Language Detection with Attention-based Deep Neural Networks*
Bushr Haddad, Zoher Orabe, Anas Al-Abood and Nada Ghneim