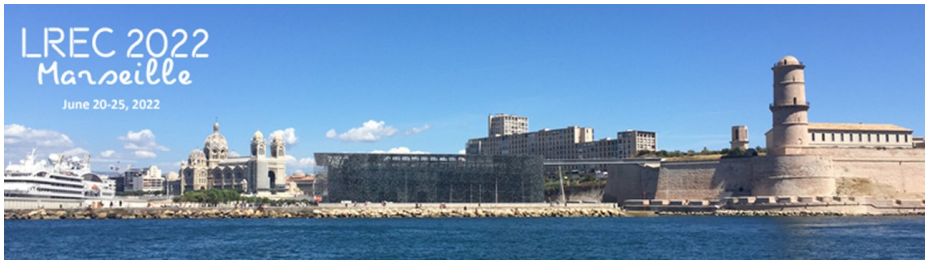

LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

The 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection

PROCEEDINGS

Editors:
Hend Al-Khalifa, Tamer Elsayed, Hamdy Mubarak,
Abdulmohsen Al-Thubaity, Walid Magdy, and Kareem Darwish

Proceedings of the LREC 2022
5th Workshop Open-Source Arabic Corpora and Processing Tools
with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection
(OSACT 2022)

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

ISBN: 979-10-95546-75-7
EAN: 9791095546757

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, third, and fourth workshops on Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) in LREC 2014, LREC 2016, LREC 2018, and LREC 2020, the fifth 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 Multilingualism and Language Technology for All, which is one of LREC 2022 hot topics. In addition to the general topics of CL, NLP and IR, the workshop gives a special emphasis on two shared tasks, namely, Qur'an QA and Fine-Grained Hate Speech Detection.

OSACT5 had an acceptance rate of 53%, where we received 15 regular papers from which 8 papers were accepted, in addition to 21 shared task papers. We believe that the accepted papers are of high quality and present a mixture of interesting topics. This year, we introduced two shared tasks: (1) the shared task on Qur'an QA 2022: Answering Questions on the Holy Qur'an, and (2) the Second Shared Task on Offensive Language and Hate Speech Detection. The Qur'an QA shared task aims to trigger state-of-the-art question answering and reading comprehension research on the Holy Qur'an. Thirty teams have registered for the task; thirteen of them submitted runs (total of 30 runs), and twelve of them eventually submitted papers for the task. The task is defined as a machine reading comprehension task on the Holy Qur'an. The participating systems are expected to provide answers to questions (posed in Modern Standard Arabic) on given passages (sets of consecutive verses) from the Holy Qur'an, where an answer is a span of text extracted from the given passage.

The other shared task aims to push the research on detecting offensive language and hate speech on Arabic Twitter in addition to determining the fine-grained hate speech type. We define offensive language as any kind of socially unaccepted language (vulgar, insults, threats, etc.). When a tweet has offensive language that targets people based on common characteristics such as race, ethnicity, ideology, gender, etc., this is considered as hate speech. We annotated data for six types of Hate Speech: Race, Religion, Ideology, Disability, Social Class, and Gender. The shared task is divided into 3 subtasks. In Subtask A ("offensive" versus "clean" tweets), 40 teams registered, and 17 teams submitted results (a total of 120 runs). In Subtask B ("hate speech" versus "no hate speech" tweets), 26 teams registered, and 12 teams submitted results (a total of 66 runs). In Subtask C ("fine-grained hate speech type"), 23 teams registered, and 10 teams submitted results (a total of 54 runs). 10 teams submitted papers describing their participation in one subtask or more, and 8 papers were accepted.

Finally, 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.

This volume documents the Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools, held on 20 June 2022 as part of the LREC 2022 conference (International Conference on Language Resources and Evaluation).

Hend Al-Khalifa, Tamer Elsayed, Hamdy Mubarak,
Abdulmohsen Al-Thubaity, Walid Magdy, and Kareem Darwish
OSACT5 Organizing Committee

Organizing Committee

Hend Al-Khalifa, King Saud University, Saudi Arabia
Tamer Elsayed, Qatar University, Qatar
Hamdy Mubarak, Qatar Computing Research Institute, Qatar
Abdulmohsen Al-Thubaity, KACST, Saudi Arabia
Walid Magdy, University of Edinburgh, UK
Kareem Darwish, aiXplain Inc., US

Program Committee

Abdelmajid Ben-Hamadou, Sfax University, Tunisia
AbdelRahim Elmadany, The University of British Columbia, Canada
Abdullah Alrajeh, King Abdulzziz City for Science and Technology, KSA
Abdulrahman Almuhareb, King Abdulzziz City for Science and Technology, KSA
Adel Alshehri, King Abdulzziz City for Science and Technology, KSA
Alexis Nasr, University of Marseille, France
Aloulou Chafik, Univeristé de Sfax, Tunisia
Areeb Alowisheq, Saudi Data and Artificial Intelligence Authority, KSA
Azzeddine Mazroui, University Mohamed I, Morocco
Bassam Haddad, University of Petra, Jordan
El Moatez Billah Nagoudi, The University of British Columbia, Canada
Fatima Haouari, Qatar University, Qatar
Fethi Bougares, Le Mans University, France
Fouzi Harrag, Ferhat Abbas University, Algeria
Hamada Nayel, Benha University, Egypt
Ibrahim Abu Farha, University of Edinburgh, Scotland
Imed Zitouni, Google, USA
Karim Bouzoubaa, Mohammad V University, Morocco
Khaled Shaalan, The British University in Dubai, UAE
Maram Hasanain, Qatar University, Qatar
Mourad Abbas, CRSTDLA, Algeria
Mucahid Kutlu, TOBB University, Turkey
Muhammad Abdul-Mageed, The university of British Columbia, Canada
Mustafa Jarrar, Bir Zeit University, Palestine
Nada Ghneim, Higher Institute for Applied Sciences and Technology, Syria
Nizar Habash, New York University Abu Dhabi, UAE
Nora Al-Twairesh, King Saud University, KSA
Omar Trigui, University of Sousse, Tunisia
Reem Suwaileh, Qatar University, Qatar
Sahar Ghannay, LIMSI, France
Sakhar Alkhereyf, King Abdulzziz City for Science and Technology, KSA
Salam Khalifa, New York University Abu Dhabi, UAE
Salima Harrat, École Normale Supérieure (Bouzaréah), Algeria
Salima mdhaffar, Le Mans University, France
Samhaa R. El-Beltagy, Newgiza University, Egypt
Saud Alashri, King Abdulzziz City for Science and Technology, KSA
Shammur Absar Chowdhury, Qatar Computing Research Institute, Qatar
Wajdi Zaghouani, Hamad Bin Khalifa University, Qatar
Waleed Alsanie, King Abdulzziz City for Science and Technology, KSA
Watheq Mansour, Qatar University, Qatar
Wissam Antoun, American University of Beirut, Lebanon
Younes Samih, Heinrich Heine Universität Düsseldorf, Germany

Table of Contents

Workshop Program

Monday 20 June 2022

Session 1: Main Workshop

9:00–9:10 *Workshop Opening*
Hend Al-Khalifa, Tamer Elsayed, Hamdy Mubarak, Abdulmohsen Al-Thubaity, Walid Magdy, and Kareem Darwish

9:10–9:50 *Keynote Talk: A proposal to accelerate innovation for Arabic Speech and Language Processing*
Hassan Sawaf, aiXplain.com

9:50–10:10 *TURJUMAN: A Public Toolkit for Neural Arabic Machine Translation*
El Moatez Billah Nagoudi, AbdelRahim Elmadany and Muhammad Abdul-Mageed

10:10–10:30 *Detecting Users Prone to Spread Fake News on Arabic Twitter*
Zien Sheikh Ali, Abdulaziz Al-Ali and Tamer Elsayed

Session 2: Main Workshop (Cont.)

11:00–11:20 *AraSAS: The Open Source Arabic Semantic Tagger*
Mahmoud El-Haj, Elvis de Souza, Nouran Khallaf, Paul Rayson and Nizar Habash

11:20–11:40 *AraNPCC: The Arabic Newspaper COVID-19 Corpus*
Abdulmohsen Al-Thubaity, Sakhar Alkhereyf and Alia O. Bahanshal

11:40–12:00 *Pre-trained Models or Feature Engineering: The Case of Dialectal Arabic*
Kathrein Abu Kwaik, Stergios Chatzikyriakidis and Simon Dobnik

12:00–12:20 *A Context-free Arabic Emoji Sentiment Lexicon (CF-Arab-ESL)*
Shatha Ali A. Hakami, Robert Hendley and Phillip Smith

12:20–12:40 *Sa'7r: A Saudi Dialect Irony Dataset*
Halah AlMazrua, Najla AlHazzani, Amaal AlDawod, Lama AlAwlaqi, Noura Al-Reshoudi, Hend Al-Khalifa and Luluh AlDhubayi

12:40–13:00 *Classifying Arabic Crisis Tweets using Data Selection and Pre-trained Language Models*
Alaa Alharbi and Mark Lee

Monday 20 June 2022 (continued)

Session 3: Qur'an QA Shared Task

14:00–14:20 *Qur'an QA 2022: Overview of The First Shared Task on Question Answering over the Holy Qur'an*
Rana Malhas, Watheq Mansour and Tamer Elsayed

14:20–14:30 *DTW at Qur'an QA 2022: Utilising Transfer Learning with Transformers for Question Answering in a Low-resource Domain*
Damith Premasiri, Tharindu Ranasinghe, Wajdi Zaghouani and Ruslan Mitkov

14:30–14:40 *eRock at Qur'an QA 2022: Contemporary Deep Neural Networks for Qur'an based Reading Comprehension Question Answers*
Esha Aftab and Muhammad Kamran Malik

14:40–14:50 *GOF at Qur'an QA 2022: Towards an Efficient Question Answering For The Holy Qu'ran In The Arabic Language Using Deep Learning-Based Approach*
Ali Mostafa and Omar Mohamed

14:50–15:00 *LARSA22 at Qur'an QA 2022: Text-to-Text Transformer for Finding Answers to Questions from Qur'an*
Youssef MELLAH, Ibtissam Touahri, Zakaria Kaddari, Zakaria Haja, Jamal Berrich and Toumi Bouchentouf

15:00–15:10 *LK2022 at Qur'an QA 2022: Simple Transformers Model for Finding Answers to Questions from Qur'an*
Abdullah Alsaleh, Saud Althabiti, Ibtisam Alshammari, Sarah Alnefaie, Sanaa Alowaidi, Alaa Alsaqer, Eric Atwell, Abdulrahman Altahhan and Mohammad Alsalka

15:10–15:20 *niksss at Qur'an QA 2022: A Heavily Optimized BERT Based Model for Answering Questions from the Holy Qu'ran*
Nikhil Singh

15:20–15:30 *QQATeam at Qur'an QA 2022: Fine-Tunning Arabic QA Models for Qur'an QA Task*
Basem Ahmed, Motaz Saad and Eshrag A. Refaee,

15:30–15:40 *SMASH at Qur'an QA 2022: Creating Better Faithful Data Splits for Low-resourced Question Answering Scenarios*
Amr Keleg and Walid Magdy

15:40–15:50 *Stars at Qur'an QA 2022: Building Automatic Extractive Question Answering Systems for the Holy Qur'an with Transformer Models and Releasing a New Dataset*
Ahmed Sleem, Eman Mohammed lotfy Elrefai, Marwa Mohammed Matar and Haq Nawaz

15:50–16:00 *TCE at Qur'an QA 2022: Arabic Language Question Answering Over Holy Qur'an Using a Post-Processed Ensemble of BERT-based Models*
Mohamemd Elkomy and Amany M. Sarhan

Monday 20 June 2022 (continued)

Session 4: Fine-Grained Hate Speech Detection Shared Task

16:30–16:40 *Overview of OSACT5 Shared Task on Arabic Offensive Language and Hate Speech Detection*
Hamdy Mubarak, Hend Al-Khalifa and Abdulmohsen Al-Thubaity

16:40–16:50 *GOF at Arabic Hate Speech 2022: Breaking The Loss Function Convention For Data-Imbalanced Arabic Offensive Text Detection*
Ali Mostafa, Omar Mohamed and Ali Ashraf

16:50–17:00 *iCompass at Arabic Hate Speech 2022: Detect Hate Speech Using QRNN and Transformers*
Mohamed Aziz Bennessir, Malek Rhouma, Hatem Haddad and Chayma Fourati

17:00–17:10 *UPV at the Arabic Hate Speech 2022 Shared Task: Offensive Language and Hate Speech Detection using Transformers and Ensemble Models*
Angel Felipe Magnossão de Paula, Paolo Rosso, Imene Bensalem and Wajdi Zaghouani

17:10–17:20 *Meta AI at Arabic Hate Speech 2022: MultiTask Learning with Self-Correction for Hate Speech Classification*
Badr AlKhamissi and Mona Diab

17:20–17:30 *CHILLAX - at Arabic Hate Speech 2022: A Hybrid Machine Learning and Transformers based Model to Detect Arabic Offensive and Hate Speech*
Kirollos Makram, Kirollos George Nessim, Malak Emad Abd-Almalak, Shady Zekry Roshdy, Seif Hesham Salem, Fady Fayek Thabet and Ensaf Hussien Mohamed

17:30–17:40 *AlexU-AIC at Arabic Hate Speech 2022: Contrast to Classify*
Ahmad Shapiro, Ayman Khalafallah and Marwan Torki

17:40–17:50 *GUCT at Arabic Hate Speech 2022: Towards a Better Isotropy for Hatespeech Detection*
Nehal Elkaref and Mervat Abu-Elkheir

17:50–18:00 *aiXplain at Arabic Hate Speech 2022: An Ensemble Based Approach to Detecting Offensive Tweets*
Salaheddin Alzubi, Thiago Castro Ferreira, Lucas Pavanelli and Mohamed Al-Badrashiny