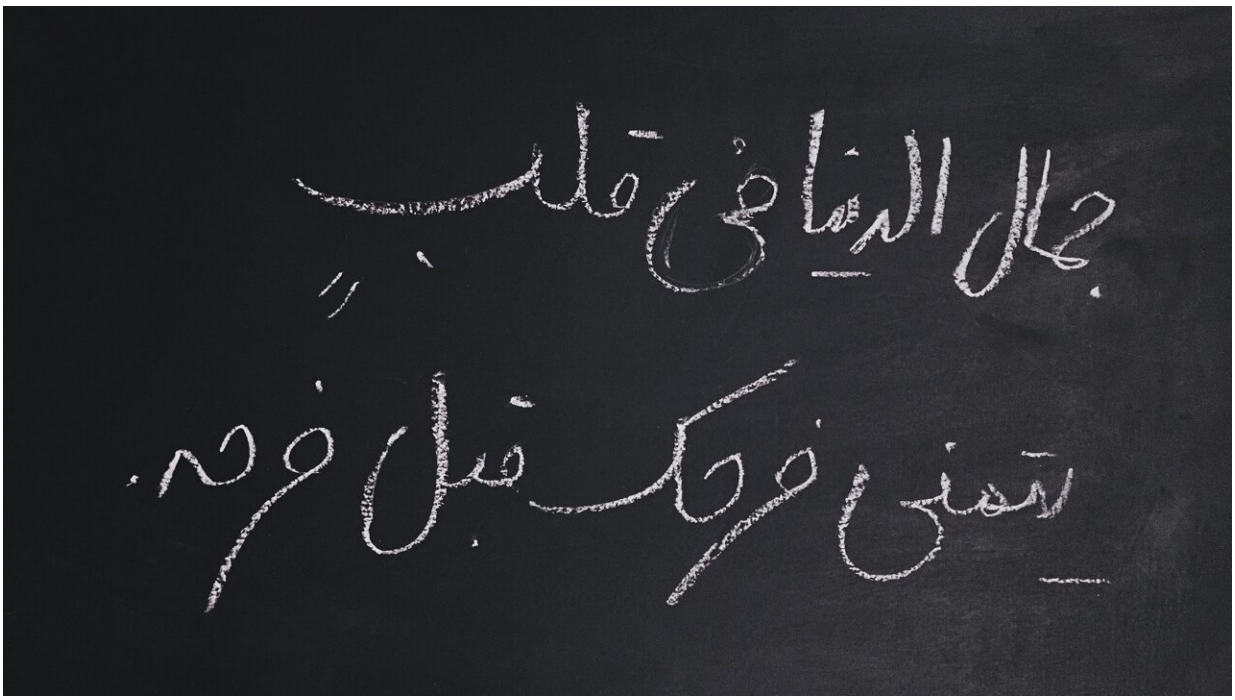
Researchers develop AI solutions for inclusion of Arabic and its dialects in Natural Language Processing

October 5 2023, by University of Sharjah



Credit: Unsplash/CC0 Public Domain

A group of researchers and engineers from the University of Sharjah have developed a deep learning system to utilize the Arabic language and its varieties in applications related to Natural Language Processing (NLP), an interdisciplinary subfield of linguistics, computer science, and

artificial intelligence.

The scientists say their project will introduce major improvements to NLP systems to accommodate the Arabic language and its dialects when programming computers to process and analyze large amounts of natural language data and assist in developing programs to enhance different language learning skills and boost translation accuracy.

The group, which includes academics and engineers, embarked on the project to assess the usability and usefulness of the Arabic language for AI-powered applications to help nearly half a billion Arabic speakers in the world to benefit from current trends in AI technologies. The results of their work have appeared in international journals.

The new AI-based system which the scientists are creating addresses the limitations NLPs encounter when processing languages other than English. The problem exacerbates with languages like Arabic whose right-to-left script and diacritics, which computers normally fail to recognize, hugely diverge from languages based on the Latin Alphabet.

To address the issue, Dr. Ashraf Elnagar, professor of computer sciences at the University of Sharjah in the United Arab Emirates, has been leading a team of academics to develop a series of computational tools that will assist programmers with the identification of not only formal Arabic but its various dialectal texts.

"The successful completion of the project has the potential to be widely adopted by the masses, as it offers numerous benefits and improvements to various AI-driven language applications and services," says Dr. Elnagar. "It has the potential to cater to a diverse range of users and industries, promoting more effective communication, accessibility, and localization."

Elaborating on the system, Dr. Elnagar says once launched, it will improve performance and user experience of applications such as machine translation, sentiment analysis, and speech recognition to accurately identify not only the standard Arabic but its numerous dialects, thereby contributing to cultural preservation, accessibility, and more effective cross-cultural communication.

Improving the status of the Arabic language with the aid of AI has become an urgent matter in Arabic speaking countries of the Middle East where computer-savvy users have started leaning on ChatGPT and other AI-driven applications to quickly generate information, execute writing assignments and improve other language skills.

Dr. Elnagar says the project draws on student research at both undergraduate and graduate levels. The project rooted in the Department of Computer Science at the University of Sharjah, showcases the remarkable talents and dedication of our students. Its inception was as a senior project by undergraduate students," notes Dr. Elnagar.

"Later, another student expanded [the] work, using it as the basis for his thesis, with a focus on textual data analysis. The project is poised to delve into the realm of audio file analysis. We take immense pride in our in-house trained students who have entirely developed this significant and impactful project."

Developers of different languages have been quick to jump on this wave of interest and currently there are numerous apps that customize for their speakers. Professor Elnagar's system will fill in a sorely missing gap as it will add Arabic, the sixth most spoken language in the world, as an operating system to AI I chatbots applications.

Developers' interest in rendering NLP-related AI tools useful to process the Arabic language and its dialects is intense. However, Dr. says his

team's system is different.

"What distinguishes our system from other AI Arabic language models is its specialized focus on detecting and processing Arabic dialects. While many models may prioritize Modern Standard Arabic or commonly spoken dialects, our system encompasses a broader range of dialectal variations.

"Developed by our in-house trained students, the technology behind our system integrates cutting-edge methodologies and deep learning techniques. Additionally, the initiative to expand its functionality from text to audio signals sets it further apart, offering a multi-modal approach to understanding and processing the Arabic language."

The team utilized a large, diverse, and bias-free dialectal dataset by merging several distinct datasets. They then trained various classical and deep learning models, including state-of-the-art Transformers, contextualizing embedding models like BERT, for region-wise and country-wise classification.

These tools can "enhance chatbot performance, which can be achieved by accurately identifying and understanding various Arabic dialects to enable chatbots to provide more personalized and relevant responses," says Professor Elnagar.

The tools can even be tailored to specific regions and cultures within the Arabic language speaking world. "This allows businesses and public services to better cater to their target audience, ensuring that the information and services provided are locally relevant and easily understood," Professor Elnagar adds.

More accurate and effective translation from and into Arabic is among the prospective outcomes of the project as the system is bound to

provide "better understanding of Arabic dialects, [help] machine translation systems [to] to provide more accurate translations, facilitating smoother communication between speakers of different dialects or languages."

Businesses and organizations are among the beneficiaries as the new AI-powered system will help them use dialect-aware sentiment analysis tools to better understand the opinions and emotions of their target audience. "This can help them tailor their marketing strategies, products, and services to cater to the specific needs and preferences of different regions or countries," said Professor Elnagar.

Asked whether external stake holders were interested in the research he and his team were conducting, Professor Elnagar said, "The project has garnered significant extracurricular interest, notably from major tech corporations like IBM and Microsoft. Additionally, Sheraa, an organization dedicated to empowering and supporting new entrepreneurs in Sharjah, has shown keen interest in the project."

"Representatives from Sheraa have engaged in discussions regarding the potential of funding the development of a commercial product based on the project's findings. This level of attention from both tech giants and entrepreneurial support entities indicates the project's potential not only as a research initiative but also as a viable commercial solution that could have broad market applications."

The AI tools the scientists are working on can likewise ensure greater accessibility for people with disabilities. "Speech recognition systems tailored to specific dialects will enable more accurate voice command recognition and transcription services for people with disabilities or those who prefer voice-based communication," said Professor Elnagar.

The project has not been without challenges, but they were successfully

addressed, notes Professor. He mentioned the issue of the lack of standardized orthography, limited resources, and labeled data, as well as the wide range of dialectal variations across Arabic-speaking regions and cultures.

Provided by University of Sharjah

Citation: Researchers develop AI solutions for inclusion of Arabic and its dialects in Natural Language Processing (2023, October 5) retrieved 17 June 2024 from https://techxplore.com/news/2023-10-ai-solutions-inclusion-arabic-dialects.html