PyArabic: A Python package for Arabic text

12 September 2022

Summary

Because text is the most common type of information representation, text processing and manipulation require recurring routines and functions. Every day, massive amounts of text are processed. Indeed, with the advent of artificial intelligence and new machine learning and deep learning enhancements, natural language processing has become a critical domain.

PyArabic is a collection of modules that provide basic functionality for manipulating Arabic texts, phrases, words, numbers, and letters. It primarily provides preprocessing tools such as normalization, tokenization, diacritics removal, number conversion, transliteration, and so on.

For years, researchers and developers who worked on machine learning algorithms for natural language processing have used the library for Arabic text preprocessing and cleaning. The library becomes more important for machine learning.

Statement of need

PyArabic is a Natural Language Processing Python package for Arabic text[1]. It is a simple library with basic functions for manipulating Arabic letters and text, such as detecting Arabic letters, Arabic letter groups and characteristics, removing diacritics, and so on. It contains the most basic and useful routines used by developers and researchers working with Arabic texts. Some key features are as follows:

- Text tokenization.
- Remove diacriticts (Harakat) from words (all, except Shadda, Tatweel, last haraka).
- Separate a word into letters and diacritics.
- Reduce diactritics of words.
- Measure tashkeel similarity (Harakats, fully or partially vocalized similarity with a template).

[1]The library can be found at [PyPi.org index](https://pypi.org/project/PyArabic/)

- Letter normalization (ligatures and Hamza).
- Numbers to words.
- Extract numerical phrases and prevocalize it.
- Unshaping texts to handle letter glyphs.
- Convert encoding and transliteration.

The PyArabic package includes five major submodules:

- Araby: Basic tools and routines for manipulating Arabic text and letters, such as tokenization and diacritics removal, are provided.
- Number: Contains routines for dealing with numbers and numeric words; allows conversion of numbers to words and words to numbers; detects numeric phrases, and more.
- Named: Provides simple tools for extracting named entities from text.
- Trans: Provides functions for converting between Arabic transliterations such as SAMPA, TIM Bukwalter, and Unicode.
- Normalize: Utility functions that are used to prepare an Arabic text for searching and indexing.

More advanced projects use PyArabic, such as:

- Adawat is an open framework for processing Arabic language that the author developed as part of his PhD research. In PhD work, we release a set of tools, the most important of which are:
- Mishkal, for restoring Arabic text diacritics (Zerrouki 2022a).
- Qalsadi is an Arabic morphology analyzer (Zerrouki 2022b).
- Tashaphyne, Arabic light stemming (Zerrouki 2022d).
- Qutrub is an Arabic verb conjugator (Zerrouki 2022c).
- The Classical Language Toolkit (CLTK)[2] (Johnson 2014) provides natural language processing support for Ancient, Classical, and Medieval Eurasia languages. CLTK integrates PyArabic functionalities for corpus importer, tokenization, text converting, and transliteration for classical Arabic (Johnson 2014), which is the form of the Arabic language used in texts from the 7th century AD to the 9th century AD (like the orthography of the Quran).

PyArabic was created to aid researchers and developers in natural language processing tasks, particularly text preprocessing. It has already appeared in several scientific publications. It is mentioned in:

- Text alignment (Mikhael 2014).
- Text classification (Abufayad 2018; Abozinadah and Jones Jr 2016; Ajlouni 2021; Habash 2021; Mgheed 2021; AlBatayha 2021).
- Sentiment analysis (Al-Horaibi and Khan 2016; Alotaibi, Mehmood, and Katib 2019; Mihi, Ali, et al. 2020; Kaibi, Nfaoui, and Satori 2019, 2020;

[2] http://cltk.org

Alharbi et al. 2020; Al-Hagery, Al-Assaf, and Al-Kharboush 2020; Oussous et al. 2020; Mihi, Ait, et al. 2020; Almutairi and Al-Hagery 2021; Mihi et al. 2022; Khabour, Al-Radaideh, and Mustafa 2022).
- Language model (Hamed, Elmahdy, and Abdennadher 2017; Alzu'bi and Duwairi 2021).
- Text preprocessing (remove diacritics, tokenization, etc.): Zhang et al. (2021)
- Lexical resources (Choe, Park, and Kim 2020)
- Text similarity (Mouty and Gazdar 2019)

PyArabic was inspired by Ar-PHP(Al-Shamaa 2022), an Arabic library for the PHP programming language that provides basic routines for web developers. Then the two libraries grow together through collaborations, and they are inspired mutually by each other. Ar-PHP provides basic routines for PHP and MySQL databases and attempts to solve web development issues such as arabic glyph rendering; however, the Ar-PHP library also includes advanced modules such as sentiment analysis, muslim prayer times, and auto-summarize (Al-Shamaa 2022).

There are many dedicated frameworks for Arabic natural language processing, like MADAMIRA(Java) (Pasha et al. 2014), FARASA(Java)(Abdelali et al. 2016), CAMeL(Python) (Obeid et al. 2020). Many multilingual frameworks, however, such as NLTK (Python) (Loper and Bird 2002), Spacy (Python) (Vasiliev 2020), and CLTK (Python) (Johnson 2014), only partially support Arabic.

In PyArabic, we focused on basic routines and build our library to be native and independent enough to be embedded in complex projects. This library was used in many projects and adopted by frameworks like CLTK(Johnson 2014), and has been inspired to build more specific libraries like TaKseem (a tokenization library for Arabic) (Alyafeai and Saeed 2020b) and Tankeeh (Arabic cleaning, normalization, and segmentation library) (Alyafeai and Saeed 2020a).

Acknowledgements

We gratefully acknowledge the contributions of Assem Chelli, Khaled Alshamaa, Lakhdar Benzahia, Mouhamad AboShokor, David Lowe, Ahmed Alq, and Arabeyes.org during the project's inception.

References

Abdelali, Ahmed, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. "Farasa: A Fast and Furious Segmenter for Arabic." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 11–16.

Abozinadah, Ehab A, and James H Jones Jr. 2016. "Improved Microblog Classification for Detecting Abusive Arabic Twitter Accounts." *International*

Journal of Data Mining & Knowledge Management Process (IJDKP) 6 (6): 17–28. https://doi.org/10.5121/ijdkp.2016.6602.

Abufayad, Tareq Issa. 2018. "Semantic Word Clustering from Large Arabic Text." PhD thesis, The Islamic University of Gaza.

Ajlouni, Moataz. 2021. "Experience Simple Transformer Library in Solving Mojaz Multi-Topic Labelling Task." In *2021 12th International Conference on Information and Communication Systems (ICICS)*, 466–67. IEEE. https://doi.org/10.1109/icics52457.2021.9464602.

Alasmari, Amal, Areej Alhothali, and Arwa Allinjawi. 2022. "Hybrid Machine Learning Approach for Arabic Medical Web Page Credibility Assessment." *Health Informatics Journal* 28 (1): 14604582211070998. https://doi.org/10.1177/14604582211070998.

AlBatayha, Duha. 2021. "Multi-Topic Labelling Classification Based on LSTM." In *2021 12th International Conference on Information and Communication Systems (ICICS)*, 471–74. https://doi.org/10.1109/ICICS52457.2021.9464531.

Al-Hagery, Mohammed Abdullah, Manar Abdullah Al-Assaf, and Faiza Mohammad Al-Kharboush. 2020. "Exploration of the Best Performance Method of Emotions Classification for Arabic Tweets." *Indonesian Journal of Electrical Engineering and Computer Science* 19 (2): 1010–20. https://doi.org/10.11591/ijeecs.v19.i2.pp1010-1020.

Alharbi, Basma, Hind Alamro, Manal Alshehri, Zuhair Khayyat, Manal Kalkatawi, Inji Ibrahim Jaber, and Xiangliang Zhang. 2020. "Asad: A Twitter-Based Benchmark Arabic Sentiment Analysis Dataset." In *KAUST Arabic Sentiment Analysis Challenge*. King Abdullah University of Sciences; Technology, Saudi Arabia.

Al-Horaibi, Lamia, and Muhammad Badruddin Khan. 2016. "Sentiment Analysis of Arabic Tweets Using Text Mining Techniques." In *First International Workshop on Pattern Recognition*, 10011:288–92. SPIE. https://doi.org/10.1117/12.2242187.

Al-Jamaan, Rawabe, Mourad Ykhlef, and Abdulrahman Alothaim. 2022. "FluSa-Tweet: A Benchmark Dataset for Influenza Detection in Saudi Arabia." In *2022 13th International Conference on Information and Communication Systems (ICICS)*, 346–51. IEEE. https://doi.org/10.1109/icics55353.2022.9811149.

Almutairi, Amjad Rasmi, and Muhammad Abdullah Al-Hagery. 2021. "Cyberbullying Detection by Sentiment Analysis of Tweets' Contents Written in Arabic in Saudi Arabia Society." *International Journal of Computer Science & Network Security* 21 (3): 112–19.

Alotaibi, Shoayee, Rashid Mehmood, and Iyad Katib. 2019. "Sentiment Analysis of Arabic Tweets in Smart Cities: A Review of Saudi Dialect." In *2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC)*, 330–35. IEEE. https://doi.org/10.1109/fmec.2019.8795331.

Alrumayyan, Nafla, and Maha Al-Yahya. 2022. "Neural Embeddings for the Elicitation of Jurisprudence Principles: The Case of Arabic Legal Texts." *Applied Sciences* 12 (9): 4188. https://doi.org/10.3390/app12094188.

Al-Sarem, Mohammad, Abdullah Alsaeedi, and Faisal Saeed. 2020. "A Deep Learning-Based Artificial Neural Network Method for Instance-Based Arabic Language Authorship Attribution." *International Journal of Advances in Soft Computing and Its Applications* 12 (2).

Al-Shamaa, Khaled. 2022. "Ar-PHP, PHP Library for Website Developers to Process Arabic Content." https://github.com/khaled-alshamaa/ar-php.

Alyafeai, Zaid, and Maged Saeed. 2020a. "Tkseem: A Preprocessing Library for Arabic." *GitHub Repository*. https://github.com/ARBML/tnkeeh; GitHub.

———. 2020b. "Tkseem: A Tokenization Library for Arabic." *GitHub Repository*. https://github.com/ARBML/tkseem; GitHub.

Alzu'bi, Dalia, and Rehab Duwairi. 2021. "Detecting Regional Arabic Dialect Based on Recurrent Neural Network." In *2021 12th International Conference on Information and Communication Systems (ICICS)*, 90–93. IEEE. https://doi.org/10.1109/icics52457.2021.9464605.

Choe, Yo Joong, Kyubyong Park, and Dongwoo Kim. 2020. "Word2word: A Collection of Bilingual Lexicons for 3,564 Language Pairs." In *Proceedings of the 12th Language Resources and Evaluation Conference*, 3036–45.

Duwairi, Rehab, Amena Hayajneh, and Muhannad Quwaider. 2021. "A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets." *Arabian Journal for Science and Engineering* 46 (4): 4001–14. https://doi.org/10.1007/s13369-021-05383-3.

Elouali, Aya, Zakaria Elberrichi, and Nadia Elouali. 2020. "Hate Speech Detection on Multilingual Twitter Using Convolutional Neural Networks." *Revue d'Intelligence Artificielle* 34 (1): 81–88. https://doi.org/10.18280/ria.340111.

Habash, Mohammad. 2021. "Team MohammadHabash at Mowjaz Multi-Topic Labelling Task." In *2021 12th International Conference on Information and Communication Systems (ICICS)*, 468–70. https://doi.org/10.1109/ICICS52457.2021.9464614.

Hamed, Injy, Mohamed Elmahdy, and Slim Abdennadher. 2017. "Building a First Language Model for Code-Switch Arabic-English." *Procedia Computer Science* 117: 208–16. https://doi.org/10.1016/j.procs.2017.10.111.

Johnson, Kyle. 2014. "CLTK: The Classical Language Toolkit." https://github.com/cltk/cltk.

Kaibi, Ibrahim, El Habib Nfaoui, and Hassan Satori. 2019. "A Comparative Evaluation of Word Embeddings Techniques for Twitter Sentiment Analysis." In *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 1–4. IEEE. https://doi.org/10.1109/wits.2019.8723864.

———. 2020. "Sentiment Analysis Approach Based on Combination of Word Embedding Techniques." In *Embedded Systems and Artificial Intelligence*, 805–13. Springer. https://doi.org/10.1007/978-981-15-0947-6_76.

Khabour, Safaa M, Qasem A Al-Radaideh, and Dheya Mustafa. 2022. "A New Ontology-Based Method for Arabic Sentiment Analysis." *Big Data and Cognitive Computing* 6 (2): 48. https://doi.org/10.3390/bdcc6020048.

Loper, Edward, and Steven Bird. 2002. "Nltk: The Natural Language Toolkit."

arXiv Preprint Cs/0205028.

Marie-Sainte, Souad Larabi. 2022. "Samee'a: A New System for Arabic Recitation Using Speech Recognition and Jaro Winkler Algorithm: Samee'a Arabic Recitation." *Kuwait Journal of Science* 49 (1).

Mgheed, Rahaf M AL. 2021. "Scalable Arabic Text Classification Using Machine Learning Model." In *2021 12th International Conference on Information and Communication Systems (ICICS)*, 483–85. IEEE. https://doi.org/10.1109/icics52457.2021.9464566.

Mihi, Soukaina, B Ait, I El, Sara Arezki, and Nabil Laachfoubi. 2020. "MSTD: Moroccan Sentiment Twitter Dataset." *International Journal of Advanced Computer Science and Applications* 11 (10): 363–72. https://doi.org/10.14569/ijacsa.2020.0111045.

Mihi, Soukaina, Brahim Ait Ben Ali, Ismail El Bazi, Sara Arezki, editor="Serrhini Laachfoubi Nabil", Carla Silva, and Sultan Aljahdali. 2020. "A Comparative Study of Feature Selection Methods for Informal Arabic." In *Innovation in Information Systems and Technologies to Support Learning Research*, 203–13. Springer International Publishing. https://doi.org/10.1007/978-3-030-36778-7_22.

Mihi, Soukaina, Brahim Ait Ben Ali, Ismail El Bazi, Sara Arezki, and Nabil Laachfoubi. 2022. "Dialectal Arabic Sentiment Analysis Based on Tree-Based Pipeline Optimization Tool." *International Journal of Electrical and Computer Engineering (IJECE)* 12 (4): 4195–205. https://doi.org/10.11591/ijece.v12i4.pp4195-4205.

Mikhael, Kamal Abou. 2014. "The Greek-Arabic New Testament Interlinear Process: Greekarabicnt. Org." *LRE-REL2*, 1.

Mouty, Rabeaa, and Achraf Gazdar. 2019. "The Effect of the Similarity Between the Two Names of Twitter Users on the Credibility of Their Publications." In *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, 196–201. IEEE. https://doi.org/10.1109/iciev.2019.8858561.

Nguyen, Khanh, and Hal Daumé. 2019. "Global Voices: Crossing Borders in Automatic News Summarization." arXiv. https://doi.org/10.48550/ARXIV.1910.00421.

Obeid, Ossama, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. "CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing." In *Proceedings of the 12th Language Resources and Evaluation Conference*, 7022–32. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.868.

Oussous, Ahmed, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih. 2020. "ASA: A Framework for Arabic Sentiment Analysis." *Journal of Information Science* 46 (4): 544–59. https://doi.org/10.1177/0165551519849516.

Pasha, Arfath, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth.

2014. “Madamira: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic.” In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1094–1101.

Solyman, Aiman, Zhenyu Wang, Qian Tao, Arafat Abdulgader Mohammed Elhag, Rui Zhang, and Zeinab Mahmoud. 2022. “Automatic Arabic Grammatical Error Correction Based on Expectation-Maximization Routing and Target-Bidirectional Agreement.” *Knowledge-Based Systems* 241: 108180. https://doi.org/10.1016/j.knosys.2022.108180.

Sun, Jimin, Hwijeen Ahn, Chan Park, Yulia Tsvetkov, and David Mortensen. 2021. “Cross-Cultural Similarity Features for Cross-Lingual Transfer Learning of Pragmatically Motivated Tasks.” In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2403–14. https://doi.org/10.18653/v1/2021.eacl-main.204.

Taha, Moaz, and Nahla Barakat. 2022. “Arabic Image Captioning: The Effect of Text Pre- Processing on the Attention Weights and the BLEU-n Scores.” *International Journal of Advanced Computer Science and Applications* 13 (August): 2022. https://doi.org/10.14569/IJACSA.2022.0130751.

Tarmom, T, E Atwell, and M Alsalka. 2019. “Non-Authentic Hadith Corpus: Design and Methodology.” In *International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2019)*.

Vasiliev, Yuli. 2020. *Natural Language Processing with Python and SpaCy: A Practical Introduction*. No Starch Press.

Yusuf, Nuhu, Mohd Amin Mohd Yunus, and Norfaradilla Wahid. 2019. “Arabic Text Stemming Using Query Expansion Method.” In *International Conference of Reliable Information and Communication Technology*, 3–11. Springer. https://doi.org/10.1007/978-3-030-33582-3_1.

Zerrouki, Taha. 2022a. “Mishkal Arabic Text Vocalization Software.” *GitHub Repository*. GitHub. https://github.com/linuxscout/mishkal.

———. 2022b. “Qalsadi Arabic Morphological Analyzer and Lemmatizer for Python.” *GitHub Repository*. GitHub. https://github.com/linuxscout/qalsadi.

———. 2022c. “Qutrub: Arabic Verb Conjugation Software.” *GitHub Repository*. GitHub. https://github.com/linuxscout/qutrub.

———. 2022d. “Tashaphyne: Arabic Light Stemmer.” *GitHub Repository*. GitHub. https://github.com/linuxscout/tashaphyne.

Zhang, Xiangliang, Qiang Yang, Somayah Albaradei, Xiaoting Lyu, Hind Alamro, Adil Salhi, Changsheng Ma, et al. 2021. “Rise and Fall of the Global Conversation and Shifting Sentiments During the COVID-19 Pandemic.” *Humanities and Social Sciences Communications* 8 (1): 1–10. https://doi.org/10.1057/s41599-021-00798-7.