Learning From Arabic Corpora
But Not Always From Arabic Speakers:
A Case Study of the Arabic Wikipedia Editions

Saied Alshahrani  Esma Wali  Jeanna Matthews
Department of Computer Science
Clarkson University, Potsdam, NY, USA
alshahsf,walie,jnm@clarkson.edu

Abstract

Wikipedia is a common source of training data for Natural Language Processing (NLP) research, especially as a source for corpora in languages other than English. However, for many downstream NLP tasks, it is important to understand the degree to which these corpora reflect representative contributions of native speakers. In particular, many entries in a given language may be translated from other languages or produced through other automated mechanisms. Language models built using corpora like Wikipedia can embed history, culture, bias, stereotypes, politics, and more, but it is important to understand whose views are actually being represented. In this paper, we present a case study focusing specifically on differences among the Arabic Wikipedia editions (Modern Standard Arabic, Egyptian, and Moroccan). In particular, we document issues in the Egyptian Arabic Wikipedia with automatic creation/generation and translation of content pages from English without human supervision. These issues could substantially affect the performance and accuracy of Large Language Models (LLMs) trained from these corpora, producing models that lack the cultural richness and meaningful representation of native speakers. Fortunately, the metadata maintained by Wikipedia provides visibility into these issues, but unfortunately, this is not the case for all corpora used to train LLMs.

1 Introduction

Natural Language Processing (NLP) is increasingly used as a key ingredient in critical decision-making systems, such as resume parsers used in sorting a list of job candidates. These NLP systems often ingest large corpora of human text, attempting to learn from past human behavior to produce systems that will make recommendations about our future world (Wali et al., 2020). The corpora of human text, which are the main ingredients in NLP systems, convey many social concepts (Cho et al., 2021), including culture, heritage, and even historic biases (Bolukbasi et al., 2016; Caliskan et al., 2017; Babaeianjelodar et al., 2020). Google News, Books Corpora, Wikipedia, and the GLUE (The General Language Understanding Evaluation) dataset (Mittermeier et al., 2021; Wang et al., 2018) are all examples of the many digital text corpora that have been used in NLP research.

Many languages are substantially under-represented in both corpus development and NLP toolchain support. For example, there are more than 7000 spoken languages around the globe, and only 300 have Wikipedia corpora. Among these 300, there is wide variation in raw corpus size as well as the ratio of articles to the number of speakers. These differences are further amplified throughout the NLP toolchain (Wali et al., 2020). Languages without large corpora also often face a lack of support in common NLP tools and unexpected errors in other tools due to a lack of testing and use. This under-represents the culture and heritage of speakers of those languages in NLP-guided decision-making.

In addition, simply having a corpus of text in a language does not necessarily represent the culture of native speakers of that language. While some corpora are originally written by native speakers, others may be written by non-native speakers or even translated from other languages (Nisioi et al., 2016). It has also been observed that some Wikipedia corpora have been developed/created through bots or automated scripts, often involving translation from other languages (Baker, 2022). This paper highlights this less discussed yet important issue of the differences between text corpora written by native speakers and those translated and generated by automated systems. We also discuss their potential effects on downstream NLP systems. As a case study, we document discrepancies between Arabic Wikipedia editions and Egyptian Arabic Wikipedia.
In Section 2, we discuss some related work, and in Section 3, we study Wikipedia and its Arabic editions, using English as a benchmark. Lastly, in Sections 4 and 5, we discuss our findings with a focus on the representativeness of NLP corpora, provide a few recommendations, and conclude with a short conclusion and pitch future work ideas.

2 Related Work

Bender et al. (2021) in an influential paper, shed light on the possible risks associated with using big data and the mitigation strategies to deal with this risk. They strongly recommend working on designing and carefully documenting datasets, as creating larger datasets and using them without having insight into their metadata could not only create documentation debt but also harm marginalized communities by introducing various kinds of biases in the results of LLMs. Without having metadata associated with the datasets, it is not possible for someone to understand training data characteristics and find ways to mitigate some of these attested issues or even unknown ones. Evaluating the approach regarding the applicability of LLMs (e.g., BERT or GPT-3) on the tasks like Natural Language Understanding (NLU) and misdirected research regarding it is another factor discussed and emphasized in this paper. Moreover, the authors advocate prioritizing LLMs’ environmental and financial costs by having their costs and resources consumed adequately reported; these costs affect the communities being least benefited by them. Lastly, a suggestion was made regarding research directions to pursue the goals of creating language technology while avoiding some of the risks and harms identified in the paper.

To help with issues related to exclusion and bias, Bender and Friedman (2018) presented the approach of including data statements in all publications and documentation for NLP systems. The approach aims to yield various short-term and long-term benefits, including unfolding how data represents the people and the world, enabling research addressing issues of bias and exclusion, promoting the development of more representative datasets, and making it convenient for researchers to consider stakeholder values as they work. Holland et al. (2020) raised the concern about the quality of data analysis methods before model development related to the cost and standardization. They presented the Dataset Nutrition Label, a diagnostic framework to aid standardized data analysis, making it more adaptable across domains. They also explored the limitations of the Label, including the challenges of generalizing across diverse datasets and guidelines for future research and policy agendas for the project. Likewise, to clarify the intended use cases of ML models and limit their usage in a context not well suited for them, Mitchell et al. (2019) suggested a framework named Model Cards to promote transparency in model reporting using short documents. Corry et al. (2021) studied dataset deprecation in ML and proposed a data deprecation framework focusing on risk, impact mitigation, appeal mechanisms, timeline, post-deprecation protocols, and publication checks that can be adapted and implemented by the ML community. They also advocate for a centralized, sustainable repository system for archiving datasets, tracking dataset deprecations, and helping to enable practices that can be integrated into research and publication processes.

To fill the gap in the standardization process in documenting datasets, Gebru et al. (2021) proposed datasheets for datasets, i.e., each dataset should be accompanied by a datasheet explaining its motivation, composition, collection process, recommended uses, etc. It aims to bridge the gap between creators and users of datasets and establish a communication channel taking a step toward ensuring transparency and accountability in datasets and ML systems. Arnold et al. (2019) proposed FactSheets to help increase trust in AI services and envisioned such documents to contain purpose, performance, safety, security, and provenance information to be completed by AI service providers for consumer examination. Denton et al. (2020) outlined a research program – a genealogy of machine learning data – for investigating how and why datasets have been created, what and whose values influence the data collection choices, and the contextual and contingent conditions of their creation. Hutchinson et al. (2021) introduced a framework for dataset development transparency that supports decision-making and accountability. The framework uses dataset development’s cyclical, infrastructural, and engineering nature to draw on best practices from the software development lifecycle.

Wikipedia is used frequently in NLP research, including multilingual NLP (Yang and Roberts, 2021; Peters et al., 2018; Devlin et al., 2018; Petroni et al., 2019; Brown et al., 2020; Wali et al., 2020; Beytía
et al., 2022; Hsu et al., 2021; Wong et al., 2021; Valentim et al., 2021; Johnson, 2020; Johnson and Lescak, 2022; Chen et al., 2021). For example, Beytía et al. (2022) documented a gender gap in Wikipedia biographical articles over a dataset of almost 6.2 million Wikipedia biographical articles across the 10 most spoken languages. The analysis was performed by proposing 4 multimodal metrics of the amount and quality of visual and written content. They found that text content favors female biographies, while the image quantity favors males, and the multilingual article coverage is biased slightly towards women. Similarly, a dataset by Valentim et al. (2021), covering 309 language editions and 33M Wikipedia articles, was presented to explore inter-language knowledge propagation by tracking the full propagation history of concepts in Wikipedia. This allows follow-up research on building predictive models with the help of aligned Wikipedia articles in a language-agnostic manner according to the concept they cover, resulting in 13M propagation instances.

Johnson and Lescak (2022) provide background about what differences might arise between different language editions of Wikipedia and how that might affect their models. The authors discuss three major ways content differences between language editions arise (local context, community and governance, and technology), recommend good practices when using multilingual and multimodal data for research and modeling, and suggest researchers expand the models available to Wikipedians for translating articles into their language.

In the space of the Arabic NLP, many researchers have studied the translation of the English language content to the Arabic language or its dialects back and forth using Machine Translation models (MTs); especially the Statistical Machine Translation models (SMTs) and the Neural Machine Translation models (NMTs), which achieved an excellent quality of translation (Al-Mannai et al., 2014; Badr et al., 2008; El-Kholy and Habash, 2010; Salloum and Habash, 2013; Sajjad et al., 2013a,b; Zbib et al., 2012). Several studies have utilized the MTs to translate the Egyptian dialect to Modern Standard Arabic (MSA) or vice versa. For example, Abo-Bakr et al. (2008) was the first work in this domain where the authors introduced a hybrid approach to translating an Egyptian sentence into its corresponding sentence in the MSA. In Mohamed et al. (2012), the author presented the opposite way, where they introduced a translator from the MSA to the Egyptian dialect. The recent work of Jeblee et al. (2014) presented many SMT systems to translate from English to Dialectal Arabic (DA) – the Egyptian Arabic dialect, using MSA as a pivot.

3 The Case of Wikipedia

Wikipedia corpora (i.e., content pages of Wikipedia) are used to train LLMs. For example, ELMo (Embeddings from Language Models) has been trained on the English Wikipedia and news crawl data (Peters et al., 2018), BERT (Bidirectional Encoder Representations from Transformers) has been trained on the BookCorpus (Zhu et al., 2015) with a crawl of the English Wikipedia (Devlin et al., 2018; Petroni et al., 2019), and GPT-3 (Generative Pre-trained Transformer) has been trained on five large datasets including the English Wikipedia as well (Brown et al., 2020).

NLP researchers find Wikipedia corpora attractive because of its large collection of multilingual content and its vast array of metadata that can be quantified and compared across the multilingual content pages (Mittermeier et al., 2021). Yet, recent works have underlined that those pre-trained LLMs embed bias, stereotypes, or even politics. Unlike many corpora, Wikipedia maintains rich metadata that allows researchers to assess the source of its contents, but little work has shown explicitly how different Wikipedia corpora impact these models (Bolukbasi et al., 2016; Caliskan et al., 2017; Yang and Roberts, 2021; Chen et al., 2021). At the same time, other recent works have also reported that the current pre-trained LLMs still under-represent the human languages despite being trained with hundreds of billions of parameters and trained on enormous datasets (Bender et al., 2021).

In the following subsections, we compare the Arabic Wikipedia editions (Modern Standard Arabic, Egyptian, and Moroccan) regarding pages to date, new pages, and top editors, besides English Wikipedia as a benchmark.1 We also specifically study the impact of problems in Egyptian Arabic Wikipedia, including large-scale auto-generation and poor translation of content pages from English.

---

1We took a data snapshot of the four Wikipedia editions’ statistics in July 2022 using the online Wikimedia Statistics service (https://stats.wikimedia.org). We contribute to the research community with our implementation of the online Wikimedia Statistics service as a Python package and command line interface. Wikistats-to-CSV (wikistats2csv) is accessible here: https://github.com/SaiedAlshahrani/Wikistats-to-CSV. See Appendix A for more details.
3.1 Arabic Wikipedia Editions

The free online encyclopedia, Wikipedia, was launched 20 years ago, in 2001, and released primarily in English (Wikipedia, 2022c). The Arabic language was one of the earliest languages added to Wikipedia. In 2004, the Arabic language content pages crossed the line of 1000 articles written by Arabic speakers to contribute to Wikipedia’s Arabic content. By 2019, Arabic content pages exceeded 1 million articles (Wikimedia Foundation, 2022b). Many Arabic Wikipedia editions appeared in the project, such as the Egyptian Arabic in 2008 and Moroccan Arabic in 2019. These are two of many dialects of the Arabic language, like Gulf Arabic, Levantine Arabic, Tunisian Arabic, and other different Arabic dialects (Habash et al., 2013).

Table 1 compares some high-level statistics of the Arabic Wikipedia editions to English Wikipedia in terms of the total number of articles (content pages), total number of pages (both content and non-content pages), total number of edits (including edits on redirects), the total number of administrators, the total number of registered users, and lastly, the total number of active users. Interestingly, Egyptian Arabic Wikipedia has a larger number of articles (content pages) than Arabic Wikipedia despite its later appearance.

3.1.1 Pages to Date

The content of Egyptian Arabic has recently grown rapidly and exponentially in the last two years. Whereas English, Arabic, and Moroccan Arabic show normal growth in their content pages (articles) over the timeline of Wikipedia.

Figure 1 shows that there were approximately 20,000 Egyptian Arabic content pages in the middle of 2019, and presently, in the middle of 2022, the Egyptian Arabic content in Wikipedia crossed the 1 million and 1/2 content articles. Almost 1.6 million content pages were created in less than 3 years, which means over 50,000 articles were created monthly, or almost 2000 pages daily. In contrast, the Arabic language content pages are currently around 1.2 million pages created in 19 years, with an average of over 5000 articles created monthly, or around 200 content pages created daily (Wikimedia Foundation, 2022b). If we associate the total number of monthly created content pages of the Egyptian Arabic Wikipedia with its latest statistics of its active users, we find that each active user would create, on average, 280 articles per month. This exponential growth of the content pages in the Egyptian Arabic Wikipedia in only 30 months is the result of the large-scale automated creation of the content pages, where one of the most active contributors confirmed this in a book (Baker, 2022); we will discuss it in detail later.

We also visualize the percentage of all page types (content and non-content) to date for the four Wikipedia editions, displaying the difference in percentage between page types to study the characteristics of each Wikipedia within itself. Figure 2 shows that all English, Arabic, and Moroccan Arabic Wikipedia have approximately 15% to 21% of content pages and approximately 79% to 85% of non-content pages of their total number of all page types. These ratios are reasonable because that is the definition of having an online free encyclopedia that aims to enable and involve people all over the globe in the creation and dissemination of knowledge. To do so effectively, users, editors, or contributors must interact with each other through talk pages, user pages, project pages, and discussion pages, generating a massive number of non-content pages in a specific Wikipedia. However, Egyptian Arabic Wikipedia opposes expected percentages, where it has approximately 20% of non-content pages and 80% of content pages, and that is a consequence of the large-scale automation of content creation.

3.1.2 New Pages

To further examine this large-scale automated creation of the content pages in the Egyptian Arabic Wikipedia and to confirm our earlier hypothesis, we studied the timeline of the three Arabic Wikipedia editions.

\[^2\]Wikipedia non-content pages include all redirects, images, categories, templates, user pages, project pages, and talk pages (Wikipedia, 2022d).
editions besides the English Wikipedia. We found that in the middle of 2020, specifically June 2020, approximately 253,000 new content pages were created in the Egyptian Arabic Wikipedia. On the other hand, nearly 23,700 new content pages were created on English Wikipedia, nearly 4,280 were created on Arabic Wikipedia, and nearly 50 on Moroccan Arabic Wikipedia, all in the same period.

Figure 3 clearly shows that the total articles (content pages) of the Egyptian Arabic Wikipedia had multiple massive spikes over the timeline of the Wikipedia project, starting from late 2019 to the beginning of 2022. Still, the most significant spike was in June 2020, when approximately 253,000 new articles (content pages) were created in one month. This is not the same as the organic creation of content pages that reflect the Egyptian people and represent their culture, beliefs, traditions, perspectives, or even dialect.

This kind of practice also appears to be inconsistent with the main purpose of the Wikipedia project; which is, according to Jimmy Wales, a co-founder of Wikipedia, "to create and distribute a free encyclopedia of the highest possible quality to every single person on the planet in their own language" (Cohen, 2008; Wikipedia, 2022a). Wikipedia should only be written, contributed to, edited, and maintained by the people. This lack of representativeness and cultural richness holds in its fold many potential problems that could impact society negatively through using deployed AI systems or NLP tools like the LLMs that have been trained on inorganic corpora (Bender et al., 2021).

3.1.3 Top Editors

Wikipedia has four types of editors: registered users (logged-in users but not in group-bot nor name-bot sets), group-bots (logged-in users who are part of a bot group), name-bots (logged-in users whose name contains ‘bot’), and anonymous users (users not logged-in but tracked by IP address) (Wikimedia Foundation, 2022c). To study the activity levels and contributions of each editor type, we visualize the percentage of all pages to date for the four Wikipedia editions by displaying the difference in percentage between editor types to study the characteristics of each Wikipedia within itself.

Figure 4 shows that Arabic and Moroccan Arabic Wikipedia editions have approximately 22% to 37% of their total number of pages created by registered users. At the same time, Egyptian Arabic...
Figure 4: The percentage of all pages to date grouped by all editor types (registered users, group-bots, name-bots, and anonymous users) for the four Wikipedia editions, displaying the difference in percentage between editor types within each Wikipedia.

Wikipedia has approximately 94% of its total pages created by registered users, and English Wikipedia has 84% of its total pages created by registered users. However, as we see in the next section, this apparent high activity level from registered users can be misleading. Important differences between English Wikipedia and Egyptian Arabic Wikipedia include the high degree of automated activity by individual registered users and the considerable gap in the total number of registered users, meaning that one registered user in Egyptian Arabic Wikipedia could create the same number of pages as hundreds or even thousands of registered users in the English Wikipedia.

3.2 Egyptian Arabic Wikipedia Problems

We investigated Egyptian Arabic Wikipedia’s top ‘registered user’ editors. We found that over 1 million articles, a surprising 63% of the total articles, in Egyptian Arabic Wikipedia have been created by one registered user called "HitomiAkane". This user has made more than 1,562,615 new creations (between articles, categories, templates, etc.), made nearly 1,615,216 edits, and created thousands of thousands of automatically generated content pages without human revision of the produced articles (Wikipedia, 2022b).

This large-scale content creation process was described by Maher Baker in his published book: *How I Wrote a Million Wikipedia Articles* (Baker, 2022). He used the English Wikipedia as a corpus and used Wikidata, which stores briefs of the articles in the form of items, each item consisting of properties and values, to generate a list of data (items) that share the same properties and values using the Wikidata Query Service (a query engine to perform queries on Wikidata database). After generating these data lists, he developed an article template where he only filled in blanks for each line of the results (data lists), which eventually became the core content of these articles. We quote the example of football player that he used to demonstrate the automation process in the book:

[label] [date of birth], [gender] is a football player from [citizenship], [gender] was born at [date of birth] in [place of birth].

The user also reported that he added the missing extra information required by Wikipedia using PHP, translated the English content to Modern Standard Arabic (MSA) using PHP’s Google Translate API, and boosted the process of creating and publishing the articles on Wikipedia using the MediaWiki Action API, a web service that allows access to a few Wiki features like page operations (create, edit, etc.) (Baker, 2022; Wikimedia Foundation, 2022a). He did not explicitly describe how he converted the MSA articles from the English translation to the Egyptian dialect. We hypothesize that the user maintained a lexicon of the most frequently used MSA words with their corresponding in the Egyptian dialect and replaced the MSA words with their Egyptian corresponding to make it look like it was produced organically by native speakers. We further suspect that many of these content articles may not have required any specific conversion to the Egyptian dialect of Arabic and thus could be considered to still be in MSA. Overall, the process used represents a relatively shallow, template-based translation of content.

According to Wikipedia’s bot policy, mass automated creation of content pages must be approved...
first, and when a user or bot operates without approval, the administrators have the right to block that user or bot (Wikipedia, 2022e). Unlike many digital corpora, Wikipedia maintains clear metadata that allows researchers to assess the source of content additions. This is an important step toward allowing researchers and users to assess whether a given corpus fits a specific use case.

Given the metadata about the Egyptian Arabic corpora, we can see that it would not be suitable corpora to learn the perspective of native speakers. Even when a Wikipedia article is a factual entry, the choice to write an article on one topic over another reflects the author’s perspective and values. Similarly, the facts chosen to add to an article vs. other possible facts not included reflect the perspective and values of the authors. It matters whether these choices are made by native speakers or by translation from other languages. We recommend that when registered users employ automated translation processes, their contributions should be marked differently than “registered user”; perhaps “registered user (automation-assisted)”.

4 Discussion
The Arabic language, in general, poses many challenges in NLP that prevent simply translating from another language like the English language due to it is morphological richness and high ambiguity (Shaalan et al., 2018; Farghaly and Shaalan, 2009). Additionally, the Arabic language has many dialectal variants, like Egyptian and Moroccan Arabic, that are different from MSA. These dialects are primarily spoken, do not have written standards, and have very few resources (Habash et al., 2013; Al-Mannai et al., 2014). Despite all these challenges the Arabic NLP faces, translating English content, especially from Wikipedia, to enrich low-resource languages’ content like Arabic using direct translation methods or tools like Google Translator. We have analyzed the Egyptian Arabic Wikipedia and found that more than 1 million articles have shallow content and are translated poorly from English to MSA. Until now, no one knows how the responsible user converted the translated MSA content to the Egyptian dialect. We suspect that most of these content articles have not truly converted to the Egyptian dialect and are still in MSA. It would be easy for users to assume that the Egyptian Arabic Wikipedia corpus was genuinely representative of the Egyptian people, their culture, heritage, or traditions. However, the many documented reasons indicate otherwise.

The other face of the lack of representativeness problem is when users or bots translate the content of the English language, for example, to other low-resource languages like Arabic using direct translation or off-the-shelf translation tools. Most of these translations done on Wikipedia content, in general, are done using direct translation, meaning that we are translating from language $A$ to language $B$. The bottleneck for this kind of translation is the quality of the translation tool. The quality of the translation is likely superior if the tool is sophis-
ticated, uses state-of-the-art technologies, and is trained on large parallel corpora of \( A \) and \( B \) languages. However, the existing off-the-shelf translation tools like Google Translator perform well, but not perfectly, and have many ethical problems like sexism and a few biases that could badly affect the translated content (Ullmann and Saunders, 2021; Lopez-Medel, 2021). It would also likely retain the sentiment, culture, and biases from the origin/source corpora rather than represent the society of native speakers of the targeted language.

Jeblee et al. (2014) designed three different translation systems: baseline MT system, where they directly translated English to Egyptian Arabic; one-step adoption MT system, where they directly translated English to MSA, used domain and dialect adoption, and translated the results to the Egyptian Arabic; and two-step adoption MT system, where they directly translated English to MSA, then used domain adoption, then in-domain MSA to dialect adoption to lastly translated the results to Egyptian Arabic. Such a complex work is what we meant by performing a sophisticated translation. We do not doubt such systems will produce a significantly accurate translation between English and Egyptian Arabic and could solve the problem of the lack of representativeness of the Egyptian Arabic Wikipedia content if it has been used. Nevertheless, the selection of which articles to write or translate and which aspects to highlight in an article would still not reflect the choices of native speakers.

As a big concern, a few researchers have studied the implications of using corpora that are automatically created, poorly translated using direct translation, automatically generated by advanced LLMs like ELMo, BERT, or GPT-3, or even the textual content of the assembled corpora using text augmentation techniques (Peters et al., 2018; Zhu et al., 2015; Brown et al., 2020; Baker, 2022; Bhattacharjee and Giner, 2022; Sahin, 2022). We believe that those LLMs, MTs, automation, and augmentation procedures will likely produce corpora full of serious issues. These corpora do not only embed bias, stereotypes, or even politics (Bolukbasi et al., 2016; Caliskan et al., 2017; Yang and Roberts, 2021; Cho et al., 2021; Chen et al., 2021), but they also do not echo the complex structure of the Arabic language and its dialects, do not express the views of the Arabic speakers, and do not represent the cultural richness and historical heritage of the Arabic language and its people.

5 Conclusion and Future Work

We studied, in this work, the Arabic Wikipedia editions (Modern Standard Arabic, Egyptian, and Moroccan) besides English Wikipedia in terms of their pages to date, new pages, and top editors, and shed light brightly on the problem of the Egyptian Arabic Wikipedia, where we found that one registered user has automated the creation of over 1 million content pages in less than 3 years and used shallow, template-based translation method that does not represent speakers of Egyptian Arabic.

We recommend that NLP practitioners avoid the inorganic, unauthentic, unrepresentative corpora in their applications (e.g., pipelines) when the goal is to learn from past human behavior and to thoroughly investigate how the corpora they do use were created, generated, or assembled; it is especially important to corpora that are produced by native speakers when the point is to examine culturally sensitive issues such as religious bias or gender bias, or political sentiment, etc. We have shown that currently, in Wikipedia, it is important to look beyond simply the “registered user” vs. “bot” distinction to recognize automated contributions, e.g., adding a “registered user (automation-assisted)” category will help us to distinguish between organically and automatically produced contributions by registered users.

In the future works, we plan to study the implications of using such unrepresentative corpora that are naively auto-created, shallowly translated, or automatically generated on the downstream applications of the NLP. We are compiling a list of alternative Egyptian Arabic corpora that have been introduced to the research community and are most likely to be organic, authentic, and representative corpora of the Egyptian Arabic dialect and its speakers. We also plan to introduce a representativeness metric that could assist in identifying the auto-generated content pages on the Wikipedia project. Lastly, we plan to design a neural network classifier that could aid in classifying the corpora in terms of representativeness.

Acknowledgments

We would like to thank the Clarkson Open Source Institute (COSI) for their support with infrastructure and hosting of our experiments. We also thank Abdullah Alshamsan and Norah Alshahrani for their help in collecting representative Arabic and Egyptian Arabic corpora for our future work.
Reproducibility


References


A Wikistats-to-CSV (wikistats2csv)

Wikistats-to-CSV (wikistats2csv) downloads Wikipedia statistics for a given Wikipedia in a CSV file format to make the online Wikimedia Statistics (Wikistats) service more accessible than it is. AI and NLP researchers and practitioners mostly use Python programming language as their first choice, and bringing this service to them as a package or command line tool saves them time and ease downloading more than one statistical CSV file in a few lines of code. We have implemented Wikistats’ three major metrics and their sub-metrics. Wikistats-to-CSV currently supports 20 queries or functions, 76 time periods, 144 filters, and 40 time intervals. We also added extra features, such as listing all Wikipedia languages with their codes, and we plan to add more features in future releases.