International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 9 Sep 2023, pp: 467-471 www.ijaem.net ISSN: 2395-5252

MoE Dataset: An Arabic Corpora of Research Center in First Karkh Director of Baghdad for the Ministry of Education

Hadeel H. Alfartosy[1, a, b)] and Hussein K. Khafaji[2, c,d)]

[1]*Iraqi Commission for Computers and Informatics, Informatics Institute for postgraduate studies, Baghdad, Iraq*

2Communication Engineering Dept Al-Rafidain University College Baghdad, Iraq

Date of Submission: 20-09-2023 Date of Acceptance: 30-09-2023

ABSTRACT: The field of Arabic language and literature exhibits significant potential for scholarly exploration and investigation. Academic scholars specializing in the field of Arabic linguistics face several obstacles, with one of the most significant being the limited accessibility to freely available Arabic corpora. This publication aims to enhance academic exploration of the Arabic language and culture. This research elucidates the challenges linked to the Arabic language, encompassing the scarcity of (1) publicly accessible Arabic corpora and (2) digitally available information in Arabic that exhibits high quality and a well-organized structure. Finally, this article presents the MOE Dataset, our most extensive dataset that is accessible to the public without charge.

I. INTRODUCTION

The Arabic language is the fifth most spoken language worldwide. It is the native tongue of approximately 422 million people and the second language of another 250 million. Arabic is categorized into three main varieties: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA). Classical Arabic (CA) includes traditional Anglican liturgical texts, Modern Standard Arabic (MSA) includes modern Anglican texts, and Dialectal Arabic (DA) consists mostly of spoken vernaculars and has no standardized written form[1].

The Arabic script encompasses a wide range of diacritics, such as I'jam (إعجام), consonant pointing, and tashkil (تشكيل), as well as extra diacritics. The second category encompasses the ḥarakat (حركات, singular haraka حركة), which are vowel marks. The term "tashkil" may be defined as the process of "forming". The primary objective of tashkil (and ḥarakat) is to serve as a phonetic guide or help in Arabic text, as the standard Arabic script alone does not offer sufficient information on accurate pronunciation. Tashkil is employed to indicate the right pronunciation of words, such as by doubling the word in pronunciation or by acting as short vowels. The term "ḥarakat", derived from the Arabic word for "motions", refers to the diacritical markings used to indicate short vowels. The diacritical marks used in Arabic script encompass Fatha, Kasra, Damma, Sukūn, Shadda, and Tanwin[2].

Despite the great use of the Arabic language, there exists a significant dearth of well-structured and high-quality digital material in Arabic. Additionally, there is a scarcity of freely accessible Arabic corpora available to the public. This work introduces MOE, an open source collection of Arabic corpora that encompasses several textual genres. These corpora hold potential for future utilization as a benchmarking resource.

II. RELATED WORKS

The utilization of corpus-based methodologies in the field of linguistics has brought forth novel perspectives in language description and diverse practical applications. These techniques enable a certain level of automated text analysis. Computer technology enables efficient and precise analysis of words, collocations, and grammatical structures in a corpus, facilitating the identification, counting, and sorting processes[3]. This significantly alleviates the laborious aspects often involved in linguistic description and substantially broadens the empirical foundation.

1. OSAC

The OSAC Arabic corpus [4] has 22,429 text fragments from a range of sources. Each text file is one of 10 types (History, Economics, Education &

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 9 Sep 2023, pp: 467-471 www.ijaem.net ISSN: 2395-5252

Family, Sports, Religious and Fatwas, Health, Low, Stories, Astronomy, and Cooking Recipes). 18 million words comprise the corpus. **Error! Reference source not found.** shows the representation of the dataset.

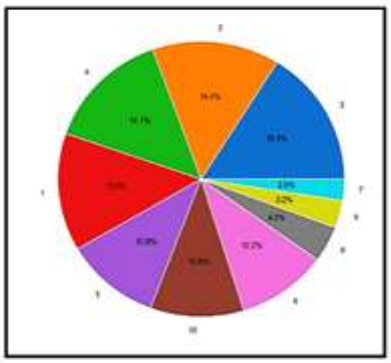

Figure 1: OSAC

2. CNN

The corpus contains 5,070 text documents from CNN Arabic's website. Each text file is one of six types: business, entertainment, Middle East news, science and technology, sports, and world news. It has 2,241,348 words. **Error! Reference source not found.** shows the representation of the dataset[5].

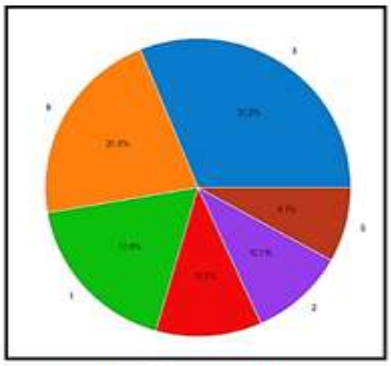

Figure 2: CNN Dataset

3. SANAD

The SANAD Dataset [1] is a massive library of Arabic news items from AlArabiya, AlKhaleej, and Akhbarona. Except for AlArabiya, all datasets have seven categories (Finance, Culture, Sports, Politics, Medicine, Religion, and Technology). We accessed 45,500 of the 190,000 SANAD documents. **Error! Reference source not found.** shows the representation of the dataset.

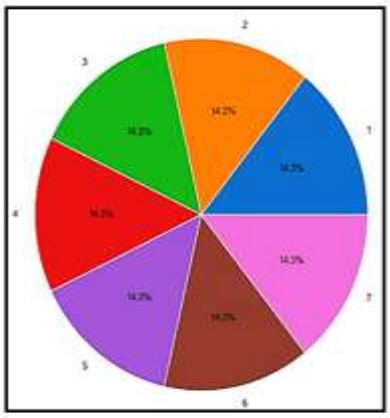

Figure 3: SANAD Dataset

4. Khaleej-2004

The Khaleej-2004 Corpus Dataset contains more than 5,000 articles which correspond to nearly 3 million words across 4 topics: International News, Local News, Economy, and Sports, in Arabic language. Containing 5,69 in HTML file format. **Error! Reference source not found.** shows the representation of the dataset.

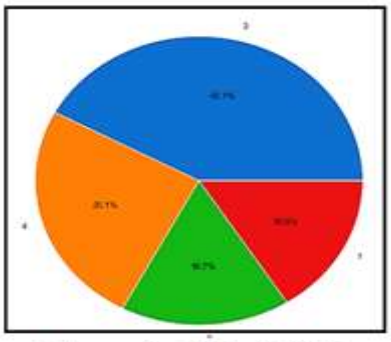

Figure 4: Kaleei-2004

III. DATA COLLECTION

It takes multiple processes to construct a labeled dataset that is representative, diversified, and accurately labeled for document categorization. The construction of a labeled dataset is a crucial stage in the development of powerful machine learning models. The dataset's quality has a direct impact on how well the classification model performs, thus each of the following processes must be given careful thought and attention. Here is a general breakdown of the procedures:

1. Determine the sources of the documents: According to the domain of the dataset to be constructed, documents can be collected from many sources, such as books, papers, journals, news, movies, websites, databases, APIs, or by crawling the web, etc.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 9 Sep 2023, pp: 467-471 www.ijaem.net ISSN: 2395-5252

2. Define Categories (Classes): Determine the classes or categories to which the documents are required to be classified. These classes could be themes, topics, sentiments, or any other relevant categorization.
3. Data Collection: It is a real step of gather a diverse set of documents that belong to the categories defined in step 2.
4. Preprocessing: Process the collected documents to prepare them for classification. This might involve tasks like removing irrelevant content (e.g., HTML tags, metadata).
5. Data Labeling: Manually assign the appropriate class labels to each document based on the defined categories. This can be a time-consuming task and might require domain expertise to ensure accurate labeling.
6. Documentation: Keep comprehensive documentation of the dataset creation process, including sources of data, preprocessing steps, labeling guidelines, and model evaluation results. This documentation is crucial for transparency and reproducibility.

This part is all about explaining how the Ministory of Education (MoE) dataset was collected. The Ministry of Education (MoE) / the Research Center in Baghdad's First Karkh District. From this source, research papers were gathered that covered a wide range of topics and fit into more than one class for the classification job. The Research Center in MoE was selected because of its stellar reputation as a comprehensive archive of Arabic scholarly works. There are numerous research papers in many disciplines at the center, making it an excellent choice for accomplishing the aims of the research.

During data collection, great effort was taken to ensure that the sample would be large enough to be representative of the whole. The objective was to cover a lot of ground in the Arabic language and a lot of different themes so that the dataset would be diverse.

Data accuracy and reliability were ensured by a stringent screening process. The selection criteria for these studies included, but were not limited to, being written in Arabic and having sufficient text for the categorization task at hand. The dataset did not include papers that did not meet these requirements. Once the study papers were chosen, they were put into different groups based on what they were about. With this multiclass method, the proposed classification system could be judged in a number of different categories within the study domain. The categorization was determined through careful content analysis and relevance assessment of the research papers in different domains.

Ethical considerations and permissions were duly addressed during the data collection process. Necessary permissions were obtained from the Research Center, and compliance with institutional or legal requirements regarding data usage and confidentiality was ensured. **Error! Reference source not found.** 5 shows the official document of permission.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 9 Sep 2023, pp: 467-471 www.ijaem.net ISSN: 2395-5252

Figure 5: Official Documents of Ethical Permissions

The collected dataset, consisting of research papers sourced from the Research Center in MoE, provides a valuable resource for training and evaluating the performance of the classification system. The inclusion of multiple classes allows for a comprehensive assessment of the system's ability to accurately classify Arabic documents across diverse topics.

Additionally, in order to ensure the robustness and generalizability of the research findings, this dissertation also incorporates a globally recognized and reliable dataset alongside the research papers collected from the Research Center. The inclusion of an external dataset provides a broader perspective and strengthens the validity of the classification models developed in this study. By incorporating both the locally sourced research papers and the globally recognized dataset, a comprehensive evaluation of the proposed classification system's performance can be achieved, ensuring the reliability and applicability of the research outcomes.

MOE dataset is collected by researcher, it has 18 classes: Administration and Economics, Literature, History, Education, Mathematics, Islamic Sciences, Philosophy, Fine Arts, Geography, Sports, Biology, Physics, Ecology, Chemistry, Psychology, Computer Science, Agriculture, and Arabic Grammar. **Error! Reference source not found.**6 shows the representation of the dataset.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 9 Sep 2023, pp: 467-471 www.ijaem.net ISSN: 2395-5252

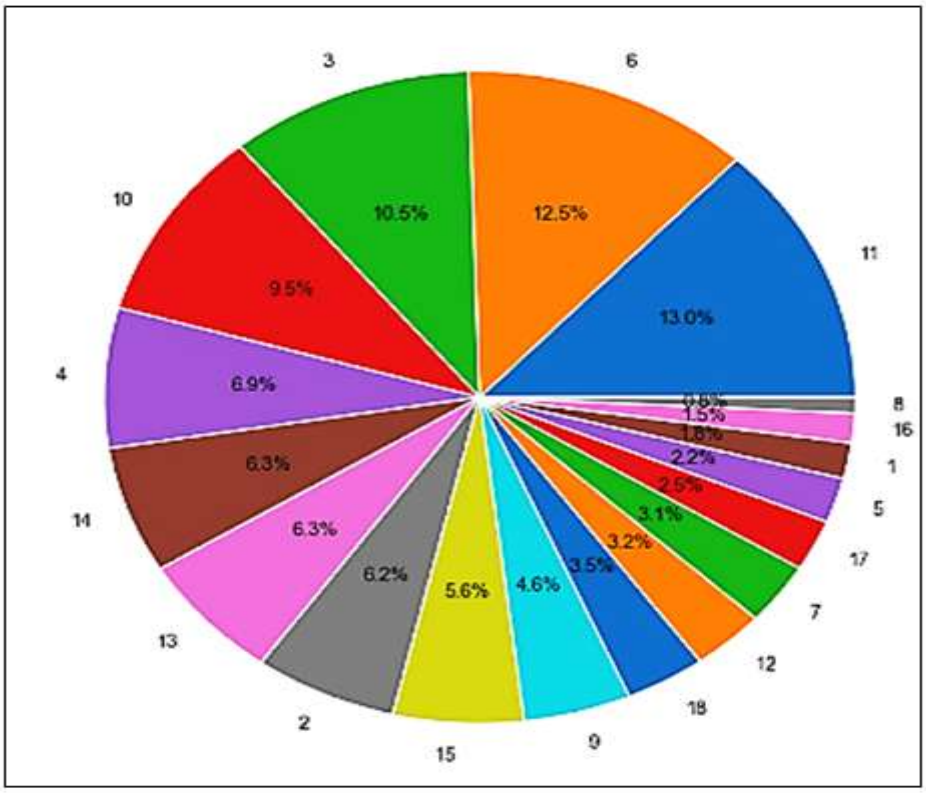

Figure 6: MOE Dataset Representation

IV. CONCLUSION

Arabic is a complex language with unique grammar and linguistic features.Unlabeled Arabic data, as in any language, refers to text that has not been annotated or categorized with specific labels or tags. This raw text can come from various sources, such as social media posts, news articles, websites, and more. Unlabeled data is a valuable resource for training machine learning models, especially for tasks like text classification.In further endeavors, our focus will be on expanding and enhancing the MOE framework. One such approach to enhance the analysis is doing thorough corpus analysis and afterwards assigning Part of Speech (POS) tags to the identified elements. Furthermore, we provide an opportunity for further academics and collaborators to expand upon the open source corpus.