Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6340–6351
Marseille, 20-25 June 2022

Masader: Metadata Sourcing for Arabic Text and Speech Data Resources

Zaid Alyafeai[1], Maraim Masoud[2], Mustafa Ghaleb[3], and Maged S. Al-shaibani[1]

[1] ICS, King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, [2] Independent Researcher
[3] Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS), KFUPM

g201080740@kfupm.edu.sa

Abstract

The NLP pipeline has evolved dramatically in the last few years. The first step in the pipeline is to find suitable annotated datasets to evaluate the tasks we are trying to solve. Unfortunately, most of the published datasets lack metadata annotations that describe their attributes. Not to mention, the absence of a public catalogue that indexes all the publicly available datasets related to specific regions or languages. When we consider low-resource dialectical languages, for example, this issue becomes more prominent. In this paper, we create *Masader*, the largest public catalogue for Arabic NLP datasets, which consists of 200 datasets annotated with 25 attributes. Furthermore, we develop a metadata annotation strategy that could be extended to other languages. We also make remarks and highlight some issues about the current status of Arabic NLP datasets and suggest recommendations to address them.

Keywords: Metadata, LR Infrastructures and Architectures, Less-Resourced

1. Introduction

The emergence of deep learning and its applications in many fields had a great impact on the development of various natural language processing (NLP) and speech techniques that were adapted to many languages. Many might correlate that to the availability of data especially with the existence of social media and the manufacturing of hardware devices that fostered research in the field, namely GPUs. Typically, we are referring to the era of deep learning which started roughly after 2010. Following that, many public Arabic NLP and speech datasets have been published in conjunction with the recent advances in deep learning (Zaghouani, 2017). Currently, there is no online centralized catalogue for Arabic NLP and speech datasets with annotated attributes. It is unclear how many online datasets there are as well as the metadata describing the datasets' characteristics, such as diversity, demographic distribution, ethical considerations, quality, and so on. This study attempts to identify the publicly available Arabic NLP datasets and to provide a catalogue of Arabic datasets to researchers. The catalogue will increase the discoverability and provide some key metadata that will help researchers identify the most suitable dataset for their research questions.

We highlight our contributions as the following:

- We create the largest catalogue with 25 attributes for 200 Arabic NLP and speech datasets.
- We design a metadata schema for annotating the datasets.
- We analyse the current status of the Arabic NLP and speech datasets, discover issues and recommend solutions.

The paper is structured as follows. Section 2 looks into previous work in the literature. Section 3 summarizes our approach to develop the catalogue. It discusses the research methodology, metadata design, and the annotation process. Section 4 outlines our findings. The results are then inspected, and issues and recommendations are highlighted in sections 5 and 6 respectively.

2. Related Work

Surveying the literature to derive analysis about a specific research field or topic is a standard practice. It helps to provide an overview on the directions and trends on subject of interest. A prominent example of such effort is (Ammar et al., 2018). They collected a large corpus of 280 million nodes. These nodes are diverse entities representing authors, papers, etc. An application of their work is the connected papers project (Eitan et al., 2020). It aims to construct a graph of related literature based on a given query. (Radev et al., 2016) extends the analysis by accounting for the citation count and extensive manual annotation on the collected literature. They curated their dataset from ACL Anthology papers. Their analysis covers various attributes including authors impact factor, h-index and collaboration. For massive analysis reports on the field of NLP, (Mohammad, 2020) surveyed the literature with 1.1 million paper information dataset collected from Google Scholar. Additionally, (Sharma et al., 2021) proposed DRIFT, a data analysis tool that presents an overview of the landscape of a queried topic. They constructed their dataset from arXiv papers' abstracts.

As the dataset collection research is vastly growing with the significant emergence of data on the web, the need to mandate such process becomes the necessity. This is, in fact, an active branch of research happening across various domains and disciplines. Consider, for instance, the systematic literature review protocol developed by (Kitchenham, 2004) that governs the data collection process for surveys. Another example is the guidelines reported by (Mbuagbaw et al., 2017) on clinical trials. There are also studies that propose standardizing the documentation of datasets. (Gebru et al., 2018) proposed datasheets for datasets. Their aim is to accompany datasets with a descriptive datasheet schema describing diverse attributes. Such attributes include operating characteristics, recommended uses, motivation, collection process and test results. Similarly, (Bender and Friedman, 2018) propose data statement, a similar standardization approach that overlaps with datasheets in some of its attributes. However, while datasheets aims to document more general information about the dataset, data statement is more specific to linguistics and NLP. Generally, there are also other studies that created metadata schema for language resources and its related processing tools. (Gavrilidou et al., 2012) presented META-SHARE which is a metadata model that describes language resources that covers datasets and processing tools. (Labropoulou et al., 2020) introduced ELG-SHARE, a rich metadata schema catering for the description of Language Resources and Technologies in addition to other entities like organizations, projects, etc. (de Jong et al., 2018) presented CLARIN which provides access to language resources and tools. They conform to the FAIR principle for findability, accessibility, interoperabililty and reusability of the data resources.

In the field of Arabic datasets, there are many studies that attempted to survey the available data resources. (Shoufan and Alameri, 2015) reviewed the NLP literature for dialectical Arabic. Their work can be considered as a quick reference to locate important contributions for certain Arabic dialects that address specific NLP features. However, this study reviewed limited literature as it was concentrated only on Arabic dialects and the research on Arabic corpora was still infant. A comprehensive approach was implemented by (Zaghouani, 2017) where he collected a list of around 80 freely available Arabic datasets including datasets that are not related to NLP. They also provide links to the datasets but some of them do not work anymore. There are also some efforts to survey specific dialects. For example, (Younes et al., 2020) provided a review of various kinds of constructed language resources (LRs) of Maghrebi Arabic dialects (MADs). They reviewed MAD raw corpora and divided it into speech corpora, speech transaction, web, and social media corpora. Recently, (Guellil et al., 2021) presented and classified 90 studies that covered classical Arabic, Modern Standard Arabic, and Arabic Dialects. Also, they provided links to around 52 NLP datasets. Another survey paper was published by (Darwish et al., 2021) to review the available tools and resources for Arabic. However, they provided links to a few datasets.

3. Methodology

The research method behind this survey follows the keyword-base literature review process (Rowley and Slack, 2004) followed by an annotation process to enrich the filtered results with metadata. Our methodology follows five steps: (i) searching resources, (ii) filtering using a selection criteria, (iii) annotating resources with metadata, (iv) validating the resources, and (v) analysing the results.

Based on a pre-selected keywords, the retrieved Arabic language resources are added to our *preliminary list* of data sources. Next, all the resources collected are filtered according to a set of inclusion criteria. The datasets that passed the filtering criteria are ported to our *final list* of datasets, and then are annotated with a set of metadata, both manually and automatically. Those which do not pass the criteria are discarded. Following that, a verification step is performed to ensure the accuracy of the metadata. Finally, the final set of resources is analysed according to the metadata and presented in this study.

The initial search and the filtering for all sources were done between July and August 2021, and the annotation process took place progressively after the data was collected and concluded on September 2021. The following subsections describe each step in more detail.

3.1. Step 1: Searching Resources

The targeted search is performed using Google Search Engine to identity Arabic NLP datasets directly. We also conduct the search against the well-known data repositories and indexing websites using a set of keywords. The selected repositories are GitHub[1], Paperswithcode[2], Huggingface[3], LREC[4], Google Scholar[5], and LDC[6]. The search combined terms related to NLP and Arabic language, such as "*NLP*", "*Natural Language Processing*" and all variations of Arabic dialects, as well as terms such as "*database*", "*dataset*", "*resource*", and "*corpus*". This step generates our *preliminary list* of data sources which consists of around 299 resources.

3.2. Step 2: Filtering with Inclusion Criteria

Our search was additionally supplemented by manually screening retrieved articles and datasets, which we perform using a set of inclusion criteria, which are as follows:

- the dataset is either in textual or spoken format.

[1] https://github.com/

[2] https://paperswithcode.com/

[3] https://huggingface.co/

[4] The International Conference on Language Resources and Evaluation http://www.lrec-conf.org/

[5] https://scholar.google.com/

[6] The Linguistic Data Consortium https://www.ldc.upenn.edu/

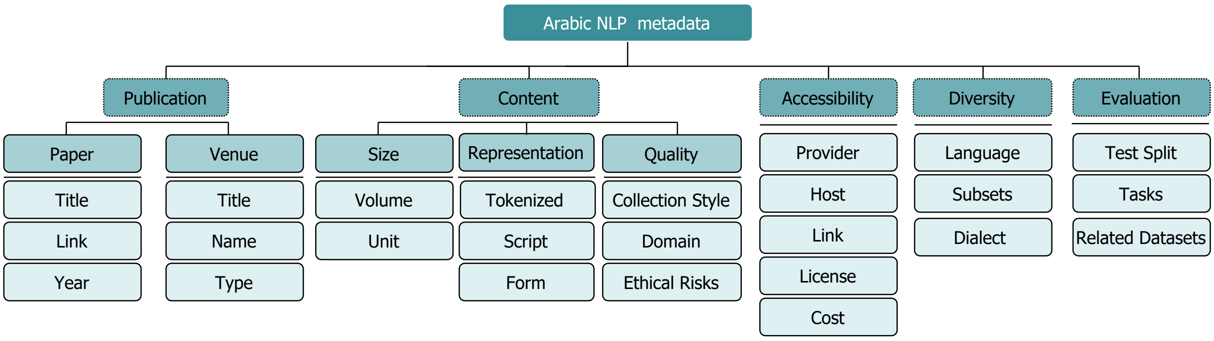

Figure 1: Metadata schema for Arabic NLP resources.

- there is a publication associated with the language resource.
- the resource is created after 2010. To limit the scope of our work and emphasize on datasets that fostered deep learning research.
- the resource, in its raw form or annotated format, is suitable for language modelling, text generation or any other NLP task.

From the *preliminary list*, we ended up with 206 papers. Our inclusion criteria removed 93 papers initially.

3.3. Step 3: The Annotation Process

By applying the inclusion criteria, the datasets search pool is reduced to the final set of considered resources. At this stage, an annotation process is applied to manually annotate them with a set of pre-agreed metadata. During this process, We consider three main goals: 1) designing metadata specific for Arabic resources, 2) deciding on the annotation format, 3) setting up an annotation workflow, and finally 4) defining an annotation task.

Metadata Selection The main motivation behind designing metadata for Arabic NLP resources is to increase their discoverability and reusability, while also representing the variety of Arabic dialects. The metadata are chosen to represent different aspects of the language resource. (Park et al., 2021)'s work serves as an example for identifying the appropriate metadata for our Arabic NLP use case. Following several revisions, the final agreed-upon metadata is represented by the taxonomy shown in Figure 1. It consists of five subcategories as follows:

- *Publication*: This subcategory concerns metadata relating to the paper, publisher and other publication details for the dataset referenced publication. It includes attributes such as the *title*, the *link*, the *year* of publication for the referenced paper, and the venue *title*, *name*, and *type*.
- *Content*: This subcategory is concerned with the content of the dataset in terms of the size, the representation and the quality. The size tag indicates the quantity of the dataset by specifying the *unit* of measurement (tokens, sentences, documents, MB, GB, TB, hours, others), as well as the number of units using the *volume* tag. The representation dimension describes the contextual information about the dataset. For example, the *tokenized* flag specifies whether the dataset is tokenized. This is useful since various tokenizers project different behaviour, and when this is not specified, it impacts downstream tasks. The *form* tag, on the other hand, defines the form of the content, being written, or spoken language, while the *script* tag describes the writing system used in the dataset (Arab,Latn,Arab-Latn,Other). The third dimension, quality, describes elements related to the data collection. It covers, for example, the *collection style* used for building the dataset (e.g crawling, translation, etc), the *ethical risk* associated with utilizing the data set (low, medium, and high), and the *domain* of the dataset (social media, etc).
- *Accessibility*: This subcategory concerns the timeliness and the reliability of access to the data. Its associated metadata includes: the name of the data *provider*, the name of the data *host*, the *link* to download the data from the host, the *licence* and the *cost* to obtain the data.
- *Diversity*: This metadata subclass is used to capture the linguistic and culture diversity within Arabic language. It covers the *language* tag to represent the language of the dataset, either being Arabic (*ar* or *multilingual* to denote a dataset that contains several languages), as well as the *subsets* tag to denote the sub-datasets that are contained inside this dataset. The last tag in this class is *dialect*. To capture the linguistic variety of Arabic, we adapted five high-level categories of dialect variations, resulting in a total of 29 dialect categories. These categories are as follows: i) MSA for Modern Standard Arabic, ii) CLS for Classic Arabic and Qura'anic text, ii) Regional dialects for the four regions (GLF, LEV, EGY, NOR)[7], iii)

[7] GLF (Gulf region), LEV (Levant region), EGY (Egypt and Sudan) and NOR (North Africa region).

country-based dialects which cover the 22 dialects spoken in Arabic-speaking countries, and finally iv) Other which includes mixed dialects and code-switched script.

- *Evaluation*: The metadata within this subcategory describes the process of using the dataset in relation to the evaluation phase of the NLP pipeline.The first, tag is the *test split*, which is deployed as a boolean flag to signal if the dataset is prepared for evaluation task by having a distinct split between the training and the test sets. The *tasks* tag defines the list of tasks to which the dataset is applied, whilst the *related datasets* attribute, lists what dataset(s) intersect with the current dataset.

Annotation Formats Based on the chosen metadata, we figured out that different annotation procedures can be applied to insert the metadata. Hence, two formats of annotation are adapted in this work; (i) manual curation, and (ii) automatic annotation of the metadata. The manual curation is performed by the human annotators via manually inspecting the dataset link, and its referenced paper. We basically use this format to extract metadata that is hard to automate or not explicitly mentioned. The second format is auto-annotation. For this format, we rely on APIs from academic publishers, such as Semantic Scholar API (Python Library)[8]. As such, most of the metadata is manually annotated, except for the publication information which was retrieved using the API.

Annotation Workflow We used Google Sheets to set up our annotation workflow, where the metadata is specified as Google Sheet columns. The datasets were annotated with a set of metadata by the four authors, who are fluent Arabic speakers and researchers in the field of natural language processing.

We define the manual annotation task as follows. For each link of Arabic language resource in our filtered pool of resources, the main goal of the task is to annotate the filtered datasets against the metadata. The annotators were instructed to follow the following set of guidelines:

1. Examine the resource link.
2. Examine the referenced paper.
3. Fill the metadata entries on Google Sheets.
4. If a particular attribute of the dataset is not mentioned, log it in the notes column.
5. If a conflict is observed between the reported metadata from the resource link and the actual published paper, mark this entry in the sheet.

3.4. Step 4: Verification of metadata

Following the completion of the annotation, a verification step is performed to confirm the accuracy of the annotations. This step was deemed necessary in order to revise the notes from the prior stage. It is done manually and involved active communication amongst the annotators. At this stage, we removed 6 extra datasets, 2 of them were duplicated datasets and 4 had wrong annotations for the year attribute (before 2010). In Figure 2, we show an example of the metadata annotations of a chosen dataset.

3.5. Step 5: Analysis

After verifying the final collection of annotated datasets, we conduct the analysis on various metadata. The findings of the analysis will be described in the next sections.

4. Findings and Representation

In this section we describe our findings in collecting all the data resources related to Arabic NLP published between 2010 and September 2021. We describe driven statistics of the datasets in addition to how we represent our data in a user friendly format.

4.1. Data Statistics

The total number of datasets included in this catalogue is 200. More than 90 % of the datasets' written format is text while the remaining is speech data. Table 1 summarizes the overall statistics of the catalogue in terms of volume. We mainly used the reported numbers in the paper and validated the numbers by downloading the dataset. If the size is different we report the numbers from the downloaded dataset. If the number can't be validated, because of the size of the dataset for example, we report the numbers from the paper. Mostly, we use tokens to represent datasets that tackle token-based tasks like named entity recognition (NER), sentences to represent datasets that are related to sentence-based tasks like sentiment analysis and documents if the size of the dataset is too large. Hours is used for speech datasets.

Table 1: Summary of the 200 Arabic NLP datasets in the Masader project in terms of volume.

Unit	**Volume**
Tokens	451,370,314
Sentences	1,236,350
Documents	51,701
Hours	3,104.1
# Datasets	200
# Datasets with Dialect Subsets	23
# Total Subsets	375

[8] https://api.semanticscholar.org/graph/v1

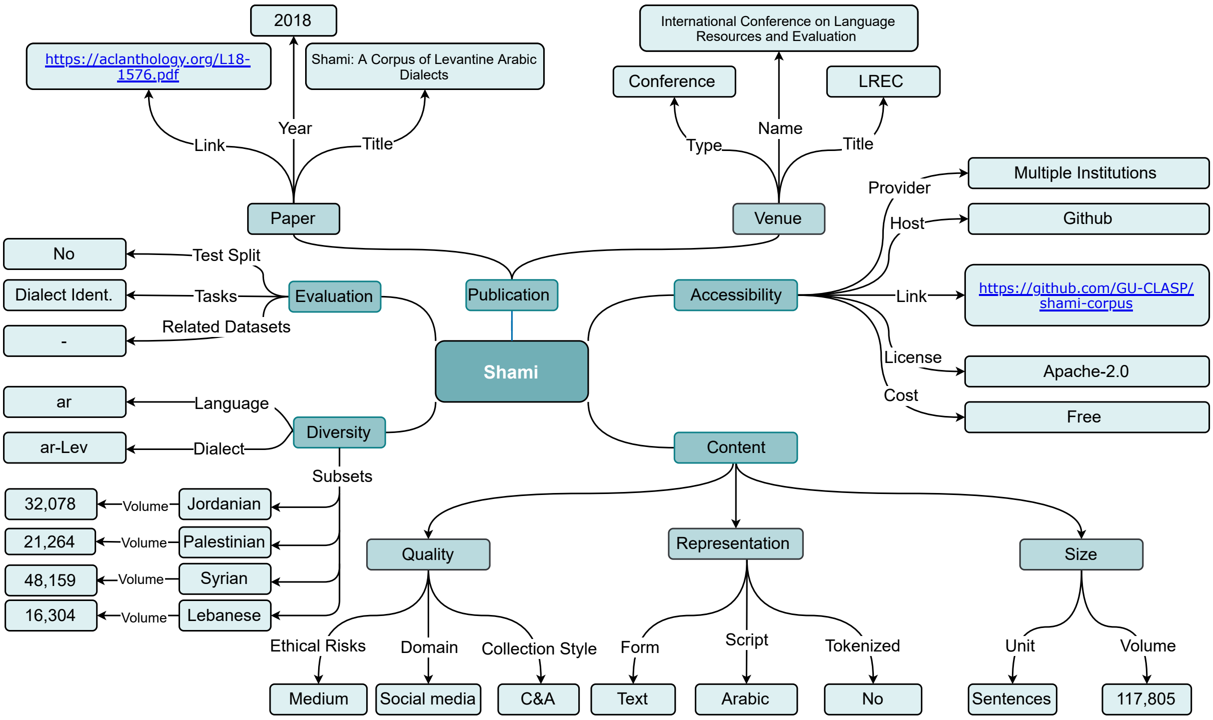

Figure 2: Example demonstrates the metadata of the Shami dataset (Abu Kwaik et al., 2018). The *subsets* tag represents the dialects and each subset (For example, Jordanian) inherits all the metadata from the superset Shami, except the *volume*.

4.2. Masader Interface

To easily navigate the sources we created a website[9] that is connected directly to the Google Sheets, allowing any updates in the sheets to be reflected immediately on the website. The website's primary interface only displays nine attributes[10]. These are deliberately chosen for piloting their relevance for academic search. The interface supports discoverability by including the following features: 1) a clickable association between the dataset and its published paper, 2) a direct link to the most recent hosted version of the dataset, 3) a clickable link on the dataset name, which leads to dataset card displaying the remaining metadata of a dataset, and finally 4) filtering and sorting based on each attribute.

5. Examining the Arabic NLP Landscape

This section provides an analysis on the surveyed datasets. We mainly focus on discussing the current trend of publishing Arabic resources and drawing some remarks about the overall status of the landscape.

5.1. Publications development

Figure 3 depicts the evolution of Arabic NLP in the light of publicly available data resources from various venues. The graph demonstrates a general growth in the number of published resources, with a particular increase in even years. This can be attributed to the large number of datasets published at the bi-annual LREC conference. We also anticipate a significant increase, particularly in 2020, with the emergence of pre-trained language models namely AraBERT (Antoun et al., 2020a), Multi-dialect BERT (Talafha et al., 2020) and Araelectra (Antoun et al., 2020b). We can also observe that most of the datasets are published in conferences and workshops.

5.2. Data Accessibility

Data accessibility is an important aspect of fostering open research. Making the data source available extends its lifespan and allows it to be utilized in the way the dataset authors intended. In our initial data sources collection, we observed that more than half of the 93 of the discarded datasets had no online presence, nor an explicit means of accessing any version of the data. Based on the examination of the data sources, there is a general trend of making the data available through open source repositories such as GitHub, GitLab and Mendeley Data. In the last three years, more than 80% of the datasets can be accessed freely on different data hosters. This trend is very promising as it shows an increased interest in making the datasets available online. In the recent years, there is a small portion of datasets that needs authentication to access either through email or registration forms. We also observed some papers that suggested contacting the corresponding author to

[9] `https://arbml.github.io/masader/`

[10] index, name, link, year, volume, unit, paper link, access, and tasks.

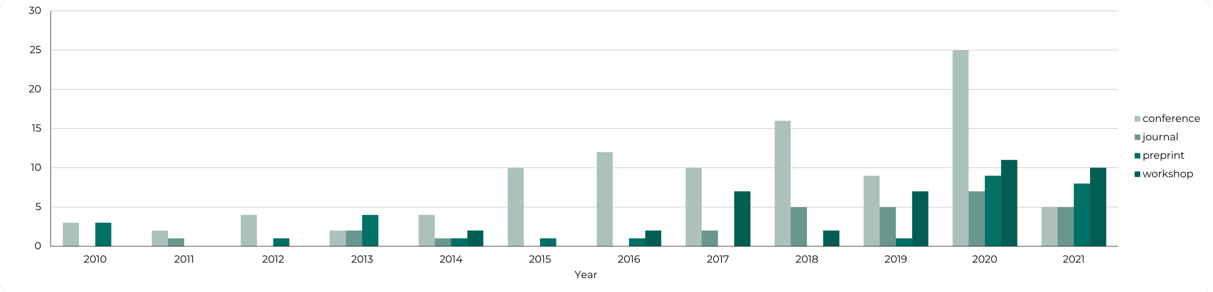

Figure 3: The count of publications across conferences, journals, preprints and workshops.

access the data privately. We didn't include such papers on our *final list*.

5.3. Data Providers and Licensing

Data providers are important to collect, annotate, distribute, and perhaps host the datasets. Another responsibility of the data provider is to select the appropriate licence for the datasets. In fact, having a proper licence is a key component of any dataset for both data providers and researchers. In terms of data providers, our findings show that the majority of the data sources we collected were created in virtue of collaboration of multiple institutions. Institutions such as QCRI, Qatar university, NYU Abu Dhabi, and Nile University are the top four providers of Arabic NLP datasets. While datasets from these prominent providers are typically accompanied by clear licensing. Unfortunately, about 50% of the datasets lack licences. Among those with explicit licence, there is a wide range of used licences. Some examples include several variations of Common Creative licences, Apache, MIT, GPL and BSD.

5.4. Dialects Diversity

As we can see from Figure 4, there were more than 20 entries out of the 200 with annotations of the dialects. These datasets are primarily intended for dialect identification tasks. The scope of the datasets we collected across the Middle East and North Africa is depicted in the Figure. The Egyptian dialect is the most prevalent, followed by Algerian, Moroccan, and Saudi dialects. Somali, Djibouti, and Mauritanian dialects are underrepresented in the surveyed datasets, with only three resources for each.

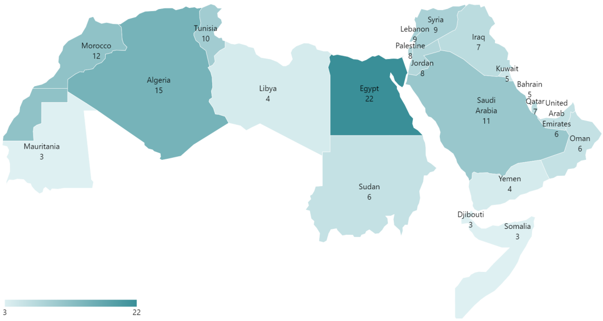

Figure 4: Dialects representation across datasets.[11]

5.5. Tasks Coverage

Figure 5 illustrates the distribution of tasks that appeared in more than one dataset. The graph shows that machine translation and sentiment analysis are the most popular tasks within Arabic NLP community. Machine translation has received an increasing attention in the literature across many languages, particularly in multilingual datasets, which explains the high frequency of publications in that area. Sentiment analysis, on the other hand, is heavily researched for a variety of reasons. Partly because the datasets for this task are primarily derived from social media sites with minimum effort, and partly because it serves as a suitable representation of everyday language that displays more sentiments. Other tasks that have presence within the community include dialect identification, topic classification, named entity recognition and speech recognition. There are also several low resource tasks that appeared at most once but are not displayed in the figure, such as poetry classification, word disambiguation, grammar checking, to name a few. Each of these sources only contains a single dataset addressing a distinct task. In contrast, datasets like KALIMAT (El-Haj and Koulali, 2013), contains annotations for multiple tasks, or evaluation suites such as ALUE which is an aggregation of multiple datasets (Seelawi et al., 2021).

6. Arabic NLP Datasets Challenges and Recommendations

This section highlights some issues of Arabic NLP datasets related to their legitimization, the haphazard collection, annotation, and the documentation practices.

Data Availability. It is encouraging that our review identified 200 datasets that were publicly available, yet discoverability seems, by all accounts, to be an issue. While a few datasets are well recognized in the field, many are not, which might potentially lead to missed research opportunities and might result in bias because of an overuse of a few potentially non-representative datasets. Further considerations in this regard, arising from our survey, include the sustainability (persistence)

[11] The Comoros Islands are not included in the map because we don't have any resources associated with it.

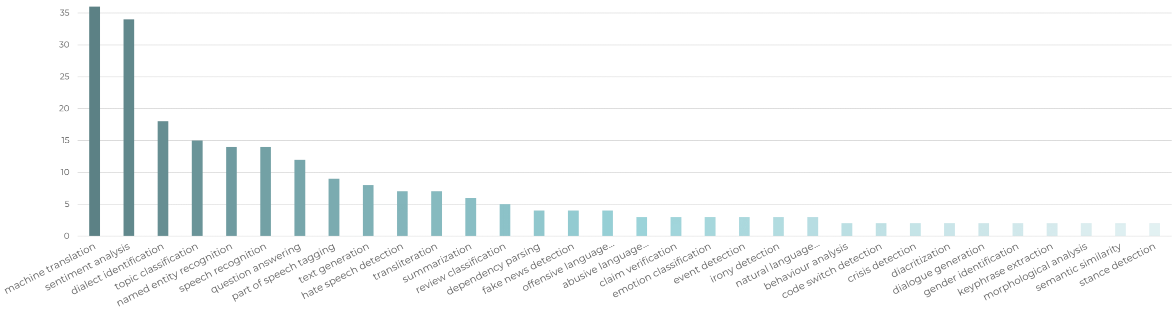

Figure 5: Tasks' histogram. We only show the tasks that appeared more than once in papers

of the dataset URLs. As there is no dedicated platform to host Arabic NLP datasets, some datasets' links appear to be inaccessible due to URLs invalidity (*orphan datasets*). We identified one obvious/clear cause of orphaned datasets, which is the termination of academic affiliation and so the broken dataset link when the dataset is published as part of the researcher's academic webpage. One possible solution to address this issue is to host the datasets on public repositories like GitHub, Gitlab, Mendeley Data, SourceForge, to name a few.

Data Documentation. Data documentation refers to the process that describes the collected data and aims at facilitating cataloguing and discoverability of the data. One key form of data documentation is metadata, which are characteristics describing the data. For a dataset to be truly reusable, adequate documentation is important to offer the necessary insights into the potential usage of the dataset. For researchers, providing such insights saves time and resources, and it suggests a reliable dataset for reuse (Perrier et al., 2020).

In this work, we analyse the documentation of Arabic NLP datasets in relations to the proposed metadata in Section 3. When we examine these datasets, we recognize that a few of them are accompanied by documentation. The majority appears to report inadequate metadata that is insufficient to make a decision on the dataset reuse. More precisely, there appears to be a pattern in which some researchers are satisfied with only publishing the direct URL link to download the dataset, or accompanying the downloadable dataset with a `README` file stating the size of the dataset, and a reference to the published paper on the dataset host page. Within the quality metadata, we observed a few instances reporting the data collection style. Another consideration includes the absence of clarity around the terms of access and use from some dataset providers. In most cases, we noticed that datasets are not accompanied by sufficient information regarding their provenance, and hence it is not possible for researchers to know if there is an appropriate ethical and governance framework underpinning the provision of these datasets. As a result of this research, we conclude that governance information, such as licencing, is a crucial part of the documentation and that, if not specified, the dataset's potential reuse may be limited. Regardless, as noted earlier, ease of access and good documentation is an important driver for researchers. Therefore, deploying a framework for documenting NLP data, such as those proposed by (Bender and Friedman, 2018) and (Gebru et al., 2018) is considered as a good step towards promoting data sharing.

Data Sharing. Data sharing is positively seen in the NLP community, with even top conferences are recognizing researchers who have shown a desire to share datasets. This process is usually volunteered, unless it is enforced internally by institutions and corporates measures. Within the Arab NLP community, we observed a high intention to support the research by sharing datasets. While some datasets are poorly documented and hardly accessible, others are well-prepared with clear documentation.

We identified some datasets that are never published, or they are inaccessible, even when there was an intention to make them available, as declared in the formal publications. In terms of openness, we also observed a pattern in which some providers require a form of registration prior to sharing the datasets. Regarding sharing dataset links, as it was detailed in the data documentation section, the unsustainability of dataset link poses a challenge, and hence data repository such as Github and Gitlab are usually adopted as a platform for sustainable data sharing.

Evaluation. NLP models are usually evaluated by training on specific tasks. In the literature, the test split is provided as an approach to evaluate models after training on the training split. In our metadata collection process, we observed that more than 60% of the datasets do not have predefined test splits. To mitigate that, researchers replicate the experiments by evaluating the old models again on a chosen random split of the data. As a result, a dataset's results will be incomparable across different NLP models.

Data Collection or Curation. Having stated annotation protocols and clear justification behind interrater agreement increases the reliability of the data. In the surveyed datasets, we have the following observations about the collection style. First, some crawling driven datasets lack any consideration for ethics and legal frameworks imposed by the platform from which the data is scraped, and the country of the data subjects. It also imposes an ethical risk by stating personal identified information. Secondly, when using machine translation to drive an Arabic version of a non-Arabic dataset, we observed missing information such as the translation models, verification process by native speakers, the reported errors, to name a few. Given the current quality of machine translation models, this approach in creating Arabic datasets opens many questions about the quality of the dataset and its potential usage. While this approach can be used to help creating datasets for some tasks, it is a risky approach if used to drive benchmark datasets for tasks such as common sense reasoning. Thirdly, as it was highlighted in previous points, not having an indication of the ethical risk of using datasets is a weak point in the Arabic NLP datasets. While each Arabic-speaking country has its own legislations and data protection acts (Abu-Ghazaleh, 2000), it is important to flag any potential risk of using the datasets for future usage.

Ethical Concerns and Privacy. Social media data, such as Twitter data, composed the greatest proportion of Arabic NLP datasets, particularly for dialect representations. Typically, such data is associated with the risk of exploiting personal information. In fact, this raises the concerns of considering data subject's right and the ethics behind using such data. Likewise, datasets acquired from publications and human-produced literature pose concerns regarding the incorporation of copyright considerations in the derived Arabic NLP datasets. In either type of the datasets, we found no explicit risk indications at any point of the NLP pipeline: collection, modelling, evaluation, or deployment. In this context, we encourage data providers to state information about the ethical risks associated with their released datasets, as well as the appropriate approach to mitigate them, in order to enrich the Arabic NLP landscape.

7. Limitations and Future Work

We recognize some limitations to our study. Firstly, given the nature of our search strategy, only datasets that are probably indexed with metadata, and whose publication contains one of our search key-phrases are likely to have been retrieved. Secondly, there exists some additional data resources available that are either with open access or regulated access (e.g LDC), but they were not explored in this study since they do not conform to our inclusion criteria. Finally, due to restricted registration or the associated fees, resources in networks of linguistic data repositories such as META-Share[12] and CLARIN were not investigated. As a future work, we plan to keep the catalogue updated by adding new datasets and also support community-based contributions where authors can submit the metadata of their datasets to our online catalogue. In addition to that we would like to focus more on speech datasets by looking into related repositories and conferences.

8. Conclusion

In this research, we created an online catalogue of 200 Arabic NLP datasets with metadata annotations. We analyzed our findings, discovered some issues and suggested some resolutions. Mainly, we recognise that the NLP field is rapidly evolving, and that both Arabic NLP researchers and practitioners recognise the value of incorporating Arabic into language technologies, particularly beyond Modern Standard Arabic. As a result, while this research provides a comprehensive analysis, it is only a snapshot in time and extra efforts are required to drive the field more in that direction.

Acknowledgements

This research was conducted under the BigScience[13] initiative for open research, a one-year-long research initiative targeting the study of large models and datasets. It was conducted as part of the data sourcing group for collecting datasets for different languages. We would like to thank the members of the data sourcing group for the insightful discussions.

9. Bibliographical References

Abu-Ghazaleh, T. (2000). *Intellectual Property Laws of the Arab Countries*. Brill, Leiden, The Netherlands.

Abu Kwaik, K., Saad, M., Chatzikyriakidis, S., and Dobnik, S. (2018). Shami: A corpus of Levantine Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., et al. (2018). Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*.

Antoun, W., Baly, F., and Hajj, H. (2020a). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Antoun, W., Baly, F., and Hajj, H. (2020b). Araelectra: Pre-training text discriminators for arabic language understanding. *arXiv preprint arXiv:2012.15516*.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., Al-Natsheh, H. T., El-Beltagy, S., Bouamor, H.,

[12] http://www.meta-share.org/

[13] https://bigscience.huggingface.co/

Bouzoubaa, K., Cavalli-Sforza, V., El-Hajj, W., Jarrar, M., and Mubarak, H. (2021). A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64:72 – 81.

de Jong, F., Maegaard, B., De Smedt, K., Fišer, D., and Van Uytvanck, D. (2018). CLARIN: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Eitan, A. T., Smolyansky, E., Harpaz, I. K., and Perets, S. (2020). Connected papers: Explore connected papers in a visual graph. Accessed: 2021-10-5.

El-Haj, M. and Koulali, R. (2013). Kalimat a multipurpose arabic corpus. In *Second workshop on Arabic corpus linguistics (WACL-2)*, pages 22–25.

Gavrilidou, M., Labropoulou, P., Desipri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V., et al. (2012). The meta-share metadata schema for the description of language resources. In *LREC*, pages 1090–1097.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., and Nouvel, D. (2021). Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Labropoulou, P., Gkirtzou, K., Gavriilidou, M., Deligiannis, M., Galanis, D., Piperidis, S., Rehm, G., Berger, M., Mapelli, V., Rigault, M., et al. (2020). Making metadata fit for next generation language technology platforms: The metadata schema of the european language grid. *arXiv preprint arXiv:2003.13236*.

Mbuagbaw, L., Aves, T., Shea, B., Jull, J., Welch, V., Taljaard, M., Yoganathan, M., Greer-Smith, R., Wells, G., and Tugwell, P. (2017). Considerations and guidance in designing equity-relevant clinical trials. *International journal for equity in health*, 16(1):1–9.

Mohammad, S. M. (2020). NLP scholar: A dataset for examining the state of NLP research. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 868–877, Marseille, France, May. European Language Resources Association.

Park, S., Moon, J., Kim, S.-D., Cho, W. I., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T. H., Lee, J., Oh, J., Lyu, S., kuk Jeong, Y., Lee, I., gyu Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A. H., Ha, J.-W., and Cho, K. (2021). Klue: Korean language understanding evaluation. *ArXiv*, abs/2105.09680.

Perrier, L., Blondal, E., and MacDonald, H. (2020). The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis. *PloS one*, 15(2):e0229182.

Radev, D. R., Joseph, M. T., Gibson, B., and Muthukrishnan, P. (2016). A bibliometric and network analysis of the field of computational linguistics. *Journal of the Association for Information Science and Technology*, 67(3):683–706.

Rowley, J. and Slack, F. (2004). Conducting a literature review. *Management research news*.

Seelawi, H., Tuffaha, I., Gzawi, M., Farhan, W., Talafha, B., Badawi, R., Sober, Z., Al-Dweik, O., Freihat, A. A., and Al-Natsheh, H. (2021). Alue: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184.

Sharma, A., Chhablani, G., Pandey, H., and Patil, R. (2021). Drift: A toolkit for diachronic analysis of scientific literature. *arXiv preprint arXiv:2107.01198*.

Shoufan, A. and Alameri, S. (2015). Natural language processing for dialectical Arabic: A survey. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 36–48, Beijing, China, July. Association for Computational Linguistics.

Talafha, B., Ali, M., Za'ter, M. E., Seelawi, H., Tuffaha, I., Samir, M., Farhan, W., and Al-Natsheh, H. T. (2020). Multi-dialect arabic bert for country-level dialect identification. *arXiv preprint arXiv:2007.05612*.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Younes, J., Souissi, E., Achour, H., and Ferchichi, A. (2020). Language resources for maghrebi arabic dialects' nlp: a survey. *Language Resources and Evaluation*, 54(4):1079–1142.

Zaghouani, W. (2017). Critical survey of the freely available arabic corpora. *CoRR*, abs/1702.07835.

A. Extra Analysis

Repositories In Figure 6, we highlight the most used repositories to host datasets. More than 50% of the datasets are hosted on GitHub. While around 23% are hosted on arbitrary websites. We notice that two main university resources are used which are QCRI (Qatar Computing Research Institute) and CAMeL (Computational Approaches to Modeling Language Lab). On the other hand most of the paid resources are on LDC (Linguistic Data Consortium). There are also other free websites for data hosting including SourceForge, GitLab and Mendeley Data. A few percentage of the datasets are also hosted in Google Drive and Dropbox.

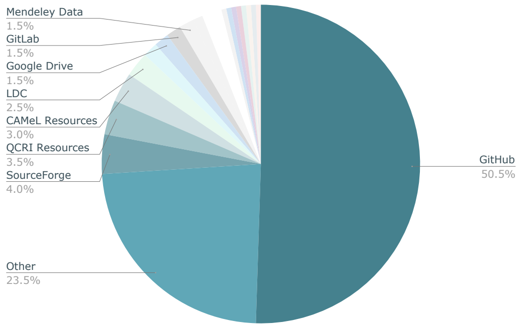

Figure 6: Most used repositories to host datasets.

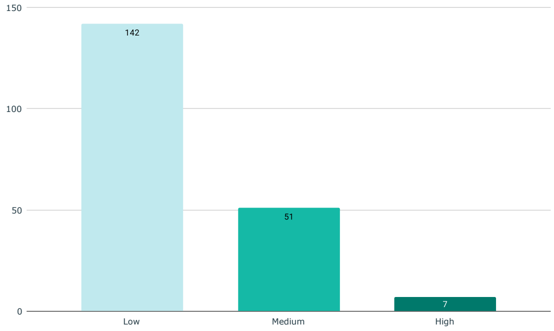

Figure 7: Distribution of datasets in terms of ethical risks.

Accessibility In Figure 8, we breakdown the three types of accessibility for datasets. We highlight that most of the datasets are free with a small percentage that either requires registration or a fee (mostly LDC). We observe a general trend of mainly publishing free data (around more than 80 % in the last few years). In the repositories that host data, we observe that more than 50 % of the data providers don't declare the type of the license for the datasets as stated in Figure 9. On the other hand around 10 % use custom licenses. Typically the most used standard licenses are variations of Creative Common, followed by Apaache, GPL and MIT.

Venues Figure 10 breakdowns the venues that are used to publish the datasets across the different years. We observe variations in the venues with around 70 unique venues across conferences, journals, workshops and preprints. As we observe from the figure the most used ones are LREC, WANLP followed by preprints (including arXiv and others). The top venues are mainly conferences and workshops.

Abstract Projection In Figures 12 and 11, we highlight all the datasets that we collected as projected embeddings. The embeddings were created by extracting the abstracts of all the datasets using Semantic Scholar API then projecting the sentence embeddings extracted from sentence-transformers[14]. The embeddings are of shape (200, 384), were projected to the 2D space using the t-SNE algorithm (Van der Maaten and Hinton, 2008). We separate the positive and negative embeddings of the x-axis for better visualization and analysis. We manually highlight 11 clusters for datasets that share common attributes. The clusters could be grouped by task (like sentiment analysis), format (like speech) or type (like parallel or multilingual datasets).

Ethical Risks Figure 7 highlights the distribution of datasets in terms of ethical risks. Datasets that could potentially contain personal information are labeled as medium. On the other hand, datasets that might contain, additionally, toxic information or hate speech text are considered as high ethical risks. All the remaining datasets are considered low. As we can observe from the Figure, most of the collected resources have low ethical risks with around 30 % having medium or high ethical risks.

Domain Representation In language modelling, it is important to include datasets that are representative not just of language diversity, but also of the domains or topics covered by the datasets. In Table 2, we highlight the various domains presented in the surveyed datasets. The majority of the datasets cover a variety of domains, because they could be scrapped from the web, Wikipedia, or collected manually. Around 30 % of the datasets are from social media, with the remaining 12 % coming from news articles. Books and reviews account for just a minor fraction of the genres in the datasets.

Table 2: Summary of domains in the surveyed datasets.

Domain	**Count**
social media	61
news articles	24
transcribed audio	16
reviews	9
books	7
other	83

[14] https://github.com/UKPLab/sentence-transformers

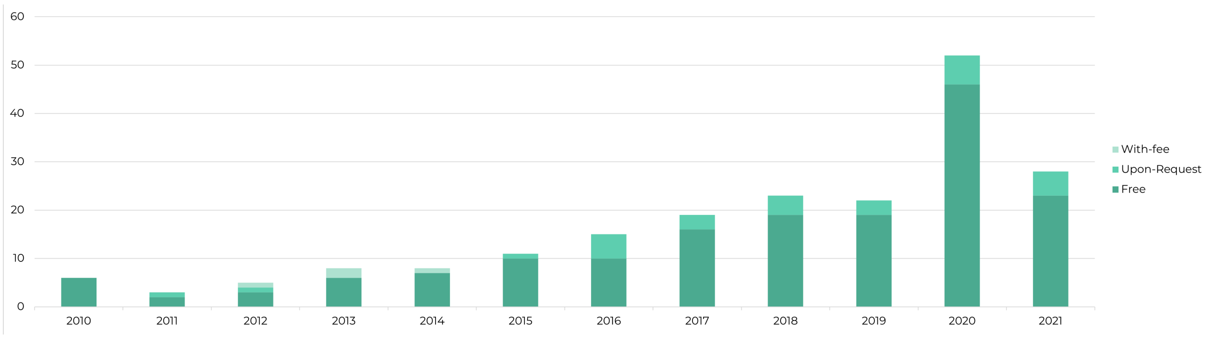

Figure 8: Breakdown of the cost associated with the datasets from 2010 to 2021.

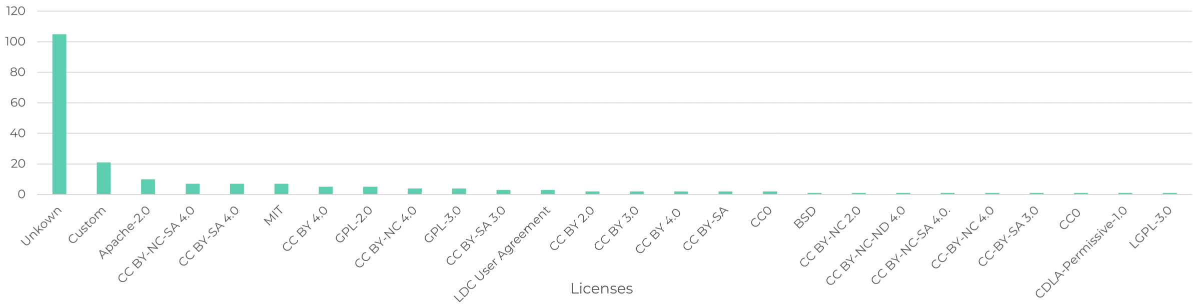

Figure 9: Distribution of licenses across the datasets.

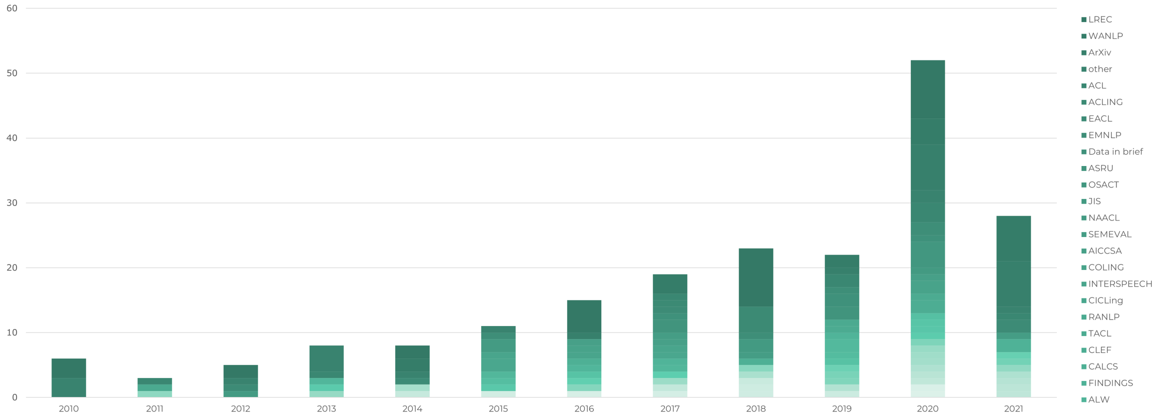

Figure 10: The count of venues within each year.

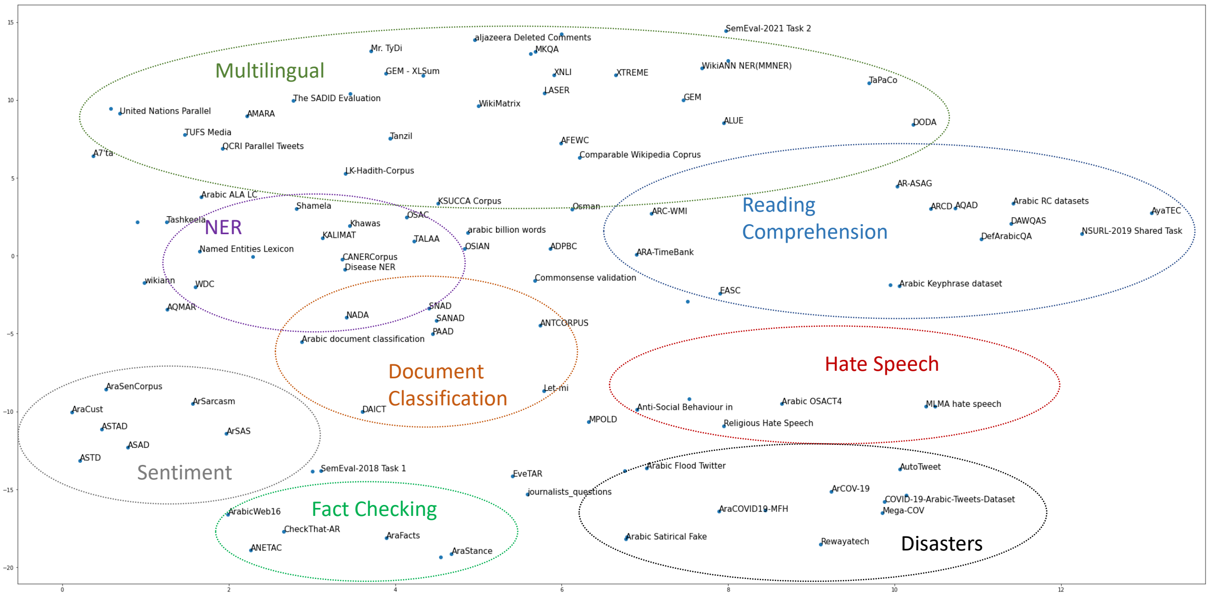

Figure 11: Positive projected embeddings of all datasets' abstracts.

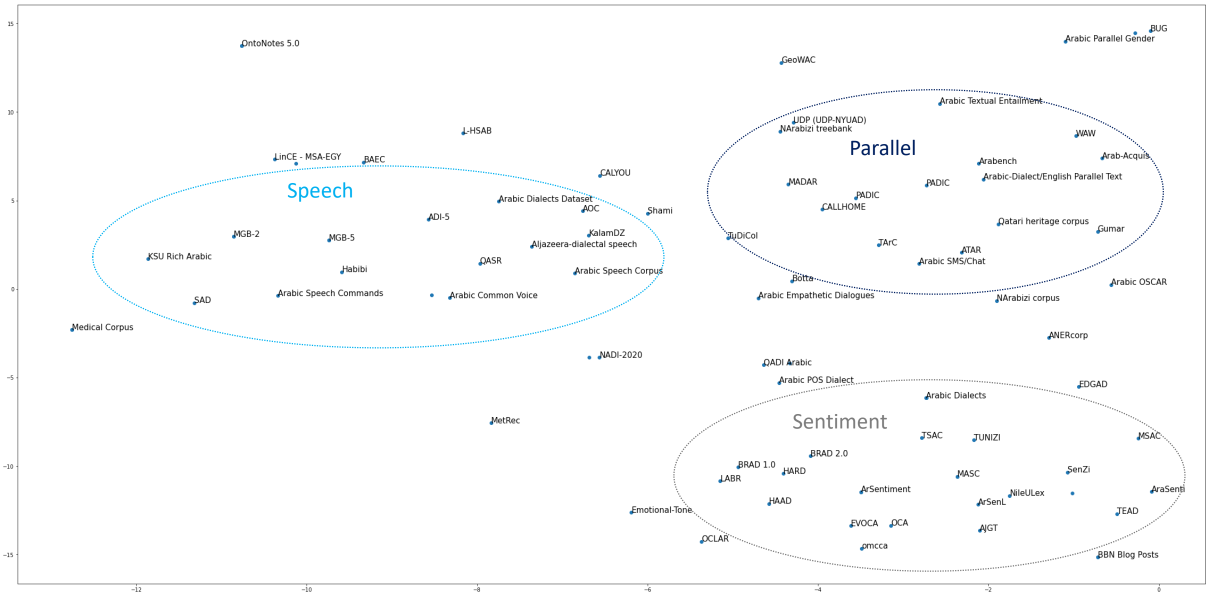

Figure 12: Negative projected embeddings of all datasets' abstracts.