**Proposed Work for Spanish and Arabic**

**Objectives**

Our overall project creates a general research platform to study civil protests, international conflict, and civil unrest using texts from English news reports. We will be expanding the development of our programs, data and services available for coding regional verbal, material conflict and cooperation methods beyond the current English-only approaches into Arabic and Spanish.

This will enable research that will advance new approaches to core questions in the social, behavioral, and economic sciences for Arabic and Spanish speaking regions. The project's data and methods will help make data-driven decisions about foreign policy, civil war prevention, human rights policies, and the effects of other factors such as environmental or economic policies in these regions.

Expanding into these languages would consist of scraping, filtering and encoding news articles from Arabic and Spanish news sites. The scraping will focus on news articles related to politics, conflict, crime and violence The encoding will be using a series of open-source natural language processing (NLP) algorithms, more specifically the BERT language representation model

The following are the basic steps in our process:

1. Scrape and filter news articles from a list of nearly 100+ Spanish news agencies and 200+ Arabic news agencies and store the results in a text database.
2. Using NLP tools, parse, lemmatize, and deploy a series of semantic and part of speech tags to the sentences in the news reports. This is done using the BERT model depending on the language of the news reports.
3. Store the results for subsequent analysis, visualization, and make them available to other researchers

We have started our work in Spanish and Arabic. Below we list all our storage requirements calculated based on the currently stored data and our projected size in a year

|  |  |  |
| --- | --- | --- |
| **Storage Name** | **Current Size** | **Projected Size in a year** |
| Spanish Dataset & Training Models | 250 GB | 1000 GB |
| Arabic Dataset & Training Models | 5 GB | 1000 GB |
| **Total** | 255 GB | 2000 GB |

Table 1: Showing storage requirement for entire project.