**Identifying, Classifying and Predicting Anti-Corruption Campaigns**

Diego Romero and Jeremy Springman

PDRI-DevLab@Penn

Our goal is to develop a system capable of predicting whether an anti-corruption campaign is “likely legitimate” or “likely fabricated.” To achieve this, we will leverage data from the Machine Learning for Peace (ML4P) Project, making our approach applicable to any country currently covered by the ML4P dataset. This database contains over 100 million newspaper articles sourced from more than 300 high-quality domestic outlets across 62 countries, classified according to their content. For our analysis, we will focus on a subset of this dataset, specifically newspaper articles categorized by ML4P as reporting on corruption, arrests, and legal action (a category that includes reporting on investigations and litigation).

We have divided our work into three main tasks. The first task involves preparing the data by classifying the articles in the ML4P subset based on several criteria: the type of corruption mentioned, whether the article reports on anti-corruption activities (e.g., arrests, investigations, or litigation), and the identity of the alleged perpetrator—categorized as a member of the opposition, a senior government official (elected or appointed), a lower-level bureaucrat, or a private citizen (e.g., a business owner or manager). This step will enable us to generate article counts reporting on anti-corruption activities at various frequencies (e.g., daily, weekly, or monthly) for each country in the dataset. Additionally, for each article about anti-corruption activity, we will have detailed information on the target of the action and the type of corruption reported—critical inputs for the final task.

We are currently working on this first task and have produced a first draft of a code book to be used to develop a training dataset. We will then use that training dataset to fine-tune a ModernBERT model to classify all newspaper articles as described. It is important to note that we will only use the title and the first three sentences of each article (which has already been translated to English), following the standard practice of the ML4P project. Once the model is trained, we will evaluate the ability of the model to accurately predict human labels on new data that the model was not trained on.

ModernBERT is a cutting-edge model well-suited for classification tasks. However, alternative state-of-the-art large language models, such as GPT, may offer better performance. That said, an important trade-off must be considered: these models require significantly more computational resources than our lab currently has access to. Pursuing this route would only be justified if the improvement in classification quality is substantial. Therefore, based on our assessment of the performance of the classification model developed in this initial task, we will either move forward with the second task or explore the feasibility of leveraging a more advanced large language model.

Second, based on the classified articles, we will employ statistical techniques to identify periods of heightened anti-corruption activity which may indicate that an anti-corruption campaign is taking place. We have not yet defined the specific statistical technique we will use.

Third, we will use information the additional information about articles which we obtained from the first task (for each article about anti-corruption activity, we will know the target of the action, and the kind of corruption being reported on), to classify the anti-corruption campaigns identified on the second task as “likely legitimate” or “likely fabricated.” We have not yet defined the specific statistical model we will use for this step.