Scientific Data

https://doi.org/10.1038/s41597-025-06334-5

Article in Press

Quantifying Global Foreign Affairs with a Multimodal Dataset of Diplomatic Websites

Received: 8 April 2025

Accepted: 18 November 2025

Cite this article as: Muğurtay, N., Şirin, K.G., Heshmat Najafabad, M. *et al.* Quantifying Global Foreign Affairs with a Multimodal Dataset of Diplomatic Websites. *Sci Data* (2025). https://doi.org/10.1038/s41597-025-06334-5

Nihat Muğurtay, Kaan Güray Şirin, Mehrdad Heshmat Najafabad, Ahmet Taha Kahya, Fazlı Göktuğ Yılmaz, Yasser Zouzou, Batuhan Bahçeci, Ayça Demir, Doğukan Tosun, Meltem Müftüler-Baç & Onur Varol

SCIENTIFIC DATA

CONFIDENTIAL
COPY OF SUBMISSION FOR PEER REVIEW ONLY

Tracking no: SDATA-25-01596B

Quantifying Global Foreign Affairs with a Multimodal Dataset of Diplomatic Websites

Authors: Nihat Muğurtay (Sabanci University), Kaan Sirin (Sabanci University), Mehrdad Heshmat Najafabad (Sabanci University), Ahmet Taha Kahya (Sabanci University), Goktug Yilmaz (Sabanci University), Yasser Zouzou (Sabanci University), Batuhan Bahceci (Sabanci University), Ayca Demir (Sabanci University), Dogukan Tosun (Sabanci University), Meltem Müftüler Baç (Sabanci University), and Onur Varol (Sabanci University)

Abstract:
This research introduces a global dataset of diplomatic news and images compiled from the official webpages of ministries of foreign affairs and chief executive offices across 156 countries spanning over 20 years. The collection provides over 1.16 million news articles and 1.18 million associated images. Our research initially shows how web scraping and Natural Language Processing (NLP) tools enhance labor-saving, novel data acquisition and processing methods. First, we extracted named entities for people, countries, and organizations mentioned in diplomatic texts. Second, GlobalDiplomacyNET processes and analyzes images published on diplomatic webpages, capturing governments' image-sharing practices. This textual and visual information together provides substantial information on countries' news-sharing habits, geographical and multilateral attention, visual assertiveness, and gender representation. GlobalDiplomacyNET is the first of its kind, offering a global corpus of textual and visual data that support novel research directions particularly in international relations and political science.

Datasets:

Repository Name	Dataset Title	Accession Number or DOI	URL to data record	Private reviewer access URL/code
Harvard Dataverse	GlobalDiplomacyNet	https://doi.org/10.7910/DVN/HYJDE0	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/HYJDE0	
GlobalDiplomacyNet-Dataset	GlobalDiplomacyNet-Dataset	https://doi.org/10.7910/DVN/HYJDE0	https://github.com/ViralLab/GlobalDiplomacyNet-Dataset/	

Quantifying Global Foreign Affairs with a Multimodal Dataset of Diplomatic Websites

Nihat Muğurtay[1,3,*], Kaan Güray Şirin[2], Mehrdad Heshmat Najafabad[2], Ahmet Taha Kahya[2], Fazlı Göktuğ Yılmaz[2], Yasser Zouzou[2], Batuhan Bahçeci[2,†], Ayça Demir[2,†], Doğukan Tosun[2,†], Meltem Müftüler-Baç[1], and Onur Varol[2,3]

[1]Faculty of Arts and Social Sciences, Sabanci University
[2]Faculty of Engineering and Natural Sciences, Sabanci University
[3]Center of Excellence in Data Analytics, Sabanci University
[*]corresponding author: Nihat Mugurtay (nihat.mugurtay@sabanciuniv.edu)
[†]these authors contributed equally to this work and are sorted alphabetically

ABSTRACT

This research introduces a global dataset of diplomatic news and images compiled from the official webpages of ministries of foreign affairs and chief executive offices across 156 countries spanning over 20 years. The collection provides over 1.16 million news articles and 1.18 million associated images. Our research initially shows how web scraping and Natural Language Processing (NLP) tools enhance labor-saving, novel data acquisition and processing methods. First, we extracted named entities for people, countries, and organizations mentioned in diplomatic texts. Second, GlobalDiplomacyNET processes and analyzes images published on diplomatic webpages, capturing governments' image-sharing practices. This textual and visual information together provides substantial information on countries' news-sharing habits, geographical and multilateral attention, visual assertiveness, and gender representation. GlobalDiplomacyNET is the first of its kind, offering a global corpus of textual and visual data that support novel research directions particularly in international relations and political science.

Background & Summary

Recent progress in novel computational techniques allows researchers to leverage large amounts of data available across disciplines and conduct studies at an unprecedented level by efficiently processing data and reducing manual effort[1]. These advanced computational techniques can be applied to in-press or online news, policy documents, social media posts, and leaders' speeches[2–9]. In this context, International Relations (IR) and Political Science (PS) are becoming significant venues for researchers to apply computational tools[10]. We contribute to these efforts by introducing **GlobalDiplomacyNET** and providing an empirical understanding of the global inventory of diplomatic news. These textual data -acquired from governmental sources- will play a critical role in helping scholars extract useful information on cutting-edge topics in global affairs. Such news often includes statements, press briefs, meetings, summits, other political events, and diplomatic reactions revealing multiple dimensions of inter-state political behavior. Beyond textual data, countries also use diplomatic images as a strategic tool[11], yet the quest for processing visual content in international relations remains underdeveloped despite its potential to reveal multiple image characteristics such as gender composition and other image attributes [12]. **GlobalDiplomacyNET** captures information from 156 countries, spanning over 20 years and contains over 1.16 million reports and 1.18 million images. Focusing on the global corpora of diplomatic news (texts and images), we demonstrate how recent advancements in computational social sciences can enhance the study of global politics and foreign policy analysis.

Our research is structured around a three-step empirical agenda. First, we acquired all textual and visual content from countries' chief executives and ministry of foreign affairs (MoFA) webpages, using automated web scraping methods. Second, GlobalDiplomacyNET uses Natural Language Processing (NLP) and computational image analysis to extract information from the data inventory for further analysis. These first two steps address the significant resource constraints traditionally associated with manual content analysis. Third, our research provides substantial information on countries' political attention and focus, uncovering how countries pay special attention to certain geographies and organizations. Using named-entities for persons, we provide a snapshot of co-occurrence networks investigating which individuals co-occur in global diplomatic news. Our image analysis reveals significant disparities in how countries prioritize gender representation.

Our research contributes to previous work, especially those projects that process large amounts of texts and capture entities, mentions, and interactions from political news and texts. Global Database of Events, Language, and Tone (GDELT)[13] has been

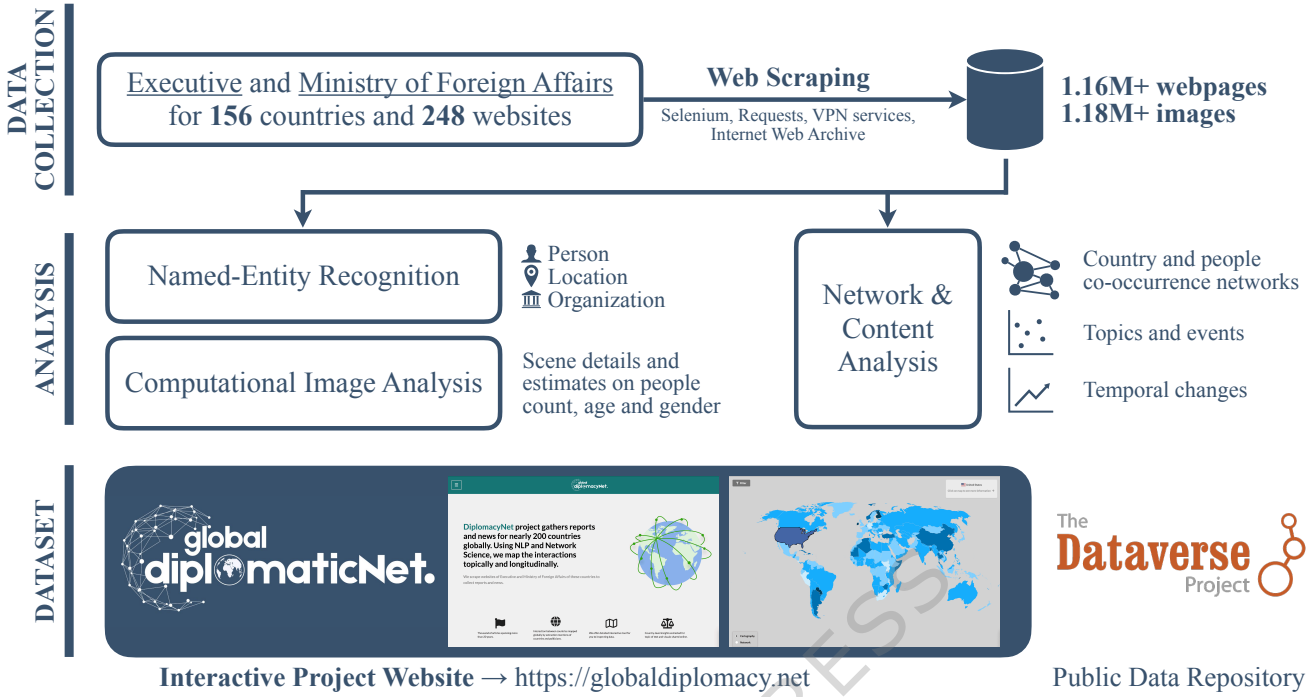


Figure 1. Schematic for the project modules. Project consist of three modules: i) data collection, ii) analysis, and iii) dataset and dissemination. We extracted content and images from 248 diplomatic websites. These dataset analyzed by using NLP and computer vision tools to conduct network and content analysis.

processing a vast amount of news articles collected from different media outlets and online sources using machine learning algorithms. Scholars use similar data to capture top-trending political issues, including migration, political discourse[14], and political violence forecasting[15]. Topic modeling also offers a significant venue for text-classification, and making it possible to extract political information[16,17], particularly on parliamentary[18] or executive speeches [19], tweets[20], news outlets, and foreign policy documents[21].

In recent years there has been an explosion in the number of texts obtained from diplomatic archives[22]. This trend has broadened the IR scholarship that analyzes large volumes of textual data. Historical diplomatic documents are processed and discussed in the realm of Foreign Policy Analysis (FPA) [23,24] regarding relationship between countries' diplomatic, political and economic behavior, with a limited number countries [25]. Acquiring and analyzing data from diplomatic archives has its own challenges [26,27]. Collecting and working with diplomatic data from open sources has traditionally entailed manual human labeling and coding. Computational techniques can overcome challenges of traditional archival research[28]. Our research demonstrates how NLP tools and LLMs can substantially reduce the human labor required to annotate political texts[29]. Further operationalization of our dataset can support these labor-intensive processes, including annotating and analyzing textual data to capture various types of information such as high-level diplomatic and economic interactions, text transcriptions, state recognition, and others[30–35]. In particular, the Correlates of War (Diplomatic Exchange)[31], DIPLOMETRICS[37,38], and DIPCON[39] datasets are notable examples of diplomatic exchange and interaction datasets. GlobalDiplomacyNET goes beyond such diplomatic datasets. For instance, while a U.S. President or Secretary of State may not visit North Korea, and diplomatic exchanges between the two countries may be absent, North Korea can still be frequently mentioned in diplomatic texts. Asymmetry between diplomatic exchanges and mentions is also another research topic that can reveal multiple dimensions of *political attention*.

When it comes to computational applications in diplomatic relations, Diplomatic Pulse[40] and GDELT emerge as promising examples of compiling global diplomatic information. However, GDELT's event-centric data does not provide full-texts directly[41]. We find that 63,318 out of 1,163,905 urls matched entries in GDELT's post-2015 data accessible through Google BigQuery. Country-level examinations further underline differences in coverage. For instance, among the 4,076 post-2015 URLs from the Turkish Ministry of Foreign Affairs (mfa.gov.tr), only 757 were present in GDELT; among the 9,728 URLs from the U.S. executive branch (whitehouse.gov), 3,938 were matched; and among the 10,910 URLs from the Russian Ministry of Foreign Affairs (mid.ru), 7,389 were found. These findings corroborate GDELT's limited diplomacy-related data coverage,

and demonstrate that coverage ratios vary considerably across countries. Diplomatic Pulse is closer to GlobalDiplomacyNET in terms of data coverage; however with GlobalDiplomacyNET, we also provide information extracted from visuals, which is one of the core components of diplomatic action and public diplomacy.

We also contribute to the existing literature by placing a growing focus on computational image analysis.[42–45]. Politicians and diplomats strategically use this visual online content to enhance their narrative in the public space[11,46]. From leaders or higher officials' stances, objects (symbols and flags), human composition (gender, and number of people) are some important aspects of visual content. For example, a previous research found that China's public diplomacy (through visuals) has a positive effect on shaping the opinion of citizens[47]. Within the scope of GlobalDiplomacyNET, we particularly focus on the gender dimension of diplomatic news, which has been a significant topic in recent years[48,49]. Despite the diverse spectrum of gender, we follow the literature and define gender as man and woman[50]. We adopt this binary framework for empirical parsimony to address recent debates.

While we have data spanning different time periods for various countries, the time series data for certain nations (such as Russia, the United States, and Japan) prove to be particularly valuable for annual dyadic aggregation. For instance, we collected diplomatic news of these countries going back to the 2000s as shown in Figure 2, extracting more information that social media data cannot offer. This is the point where researchers can integrate our data with existing dyadic datasets for different domain-specific purposes. Examples of such inferential uses can include political similarity (e.g. UNGA Voting Similarity[51] , trade partnership, under-reported global finance datasets[42], leader visits[34,52], interstate disputes [36], governance indicators[53–55], global surveys (e.g., Gallup World Poll), conflict intensity[56,57]. However, we can point research directions where social media is an important medium for sharing diplomatic news and leaders' views, leading a term *Twiplomacy*. In addition to official websites, the current literature also suggests leaders' social media presence is effective and visible for diplomacy [58–60]. Censorship and shadow banning also leads politicians to construct their own online spaces and deplatforming of their followers like Trump's Truth Social platform. These fringe communities can serve close followers, but their diplomacy related messages can have a different tone and framing. Zhang *et al.* compare partisan asymmetry for Trump's activity on Twitter and Truth Social [61]. Cross-platform analysis between social media and official channels can offer interesting insights and our dataset can be beneficial to conduct such analysis.

Methods

GlobalDiplomacyNETproject consists of several modules for data collection and analysis. In this section, we present the data collection methodologies and the approaches used to extract information from the gathered content. In Figure 1, we present the schematic of the project to highlight the key components and their interactions with each other.

Curation of country list and institutions

GlobalDiplomacyNET relies on the traditional (state-centric) definition of diplomacy, namely the states' political activity conducted by official authorities who oversee and implement foreign policy [62,63]. GlobalDiplomacyNET's country sample is composed of UN Member sovereign states, since a country's diplomatic salience is affected by its recognition status. In other words, we excluded de-facto states such as South Ossetia and the Turkish Republic of Northern Cyprus (TRNC). However, we included the West Bank and Gaza Strip (mofa.pna.ps) and Taiwan (en.mofa.gov.tw) due to their high salience in the current geopolitical agenda. Particularly, recognition of Palestine as a sovereign entity have been one of the acute topics in international politics. We also excluded the British Overseas Territories due to their limited sovereignty.

Foreign policy implementation and diplomatic professions are primarily handled by countries' ministries of foreign affairs under the supervision of the heads of government (chief executives). This definition also defines the principal actors of diplomacy as presidents, prime ministers, and foreign ministers[64]. We curated and compiled the GlobalDiplomacyNET dataset from countries' ministry of foreign affairs and chief executive webpages. Considering the total of 156 countries and their 248 distinct websites, we encountered some exceptions in terms of their organizations. In case MoFA news focuses solely on foreign ministry activities, we collected data from both MoFA and Chief Executive websites. Some countries such as Liberia (mofa.gov.lr) and Maldives (foreign.gov.mv) had gaps in their news coverage. Observed temporal gaps can occur due to leadership changes, civil conflict and some external shocks such as COVID-19. We use Wayback Machine to fill those gaps when there is a *systematic* deletion of news. As a limitation, we were unable to find some conflict-affected and small states' official websites during the phase of scraping process. In addition, some webpages did not respond during the scraping phase. In total, 40 UN Members are excluded in the current version of our dataset. In some other examples, governments' official websites announce all executive and MoFA activities (e.g., Ireland, https://www.gov.ie/), filtering particular institutions or agencies. For semi-presidential systems, we scraped predominantly the president as the primary diplomatic actor. Since premier-presidential and president-parliamentary systems (e.g. Russia, http://en.kremlin.ru/) put a level of complexity, we only focused on president in such systems for conceptual and empirical parsimony. Prime-ministers in semi-presidential systems generally have a limited diplomatic salience when compared to president.

The number of websites exceeds the number of unique countries because we collected data from both MoFA and executive offices. For some countries, such as the United States, each presidential administration maintains separate archived webpages for both the chief executive and the secretary of state. During the research phase, we strictly complied with legal provisions within the scope of open-science practices [65]. In other words, GlobalDiplomacyNET does not include any personal data, and all data come from publicly available sources.[66].

Collecting information from diplomatic websites by web scraping

Curation of the country list and the corresponding diplomatic websites present distinct challenges in collecting news content. These challenges emerge due to the technologies used in web development, completeness of historical records, or the anti-scraping methods employed to reduce automated traffic to the websites.

For this project, we have to develop distinct web scrapers for each domain, since all websites have a different structure. First, we identified the pages where the diplomatic news is listed either by some filters such as time and topic. Our scraping scripts collect URLs for news articles by visiting different pages or dynamically loading the next batch of articles by triggering JavaScript events. For static webpages, the use of standard web scraping tools like Python's `requests` package was sufficient to iterate over the pages or send requests to API endpoints of the websites. Some dynamic websites require user interactions by clicking certain buttons or scrolling down the pages. In such scenarios, we implement emulators using `Selenium` library to mimic user behaviors. Once the URLs were collected, we visited and collected content from the websites as our second step.

Since some websites contain thousands of articles, platforms implement measures to prevent scraping such as CAPTCHAs to test authentic behaviors and protect sites from high traffic by using reverse proxy services like CloudFlare. On the other hand, we develop our custom scrapers to respond to such measures and send our requests from different IP addresses by using commercial VPN services. We noticed some government webpages restrict access or may not respond from particular countries' IP addresses. Another challenge we face is the removal of content or websites altogether. When countries make changes to implement extra layers of protection or change their design, their content may become unavailable and custom scrapers also need reimplementation. For instance, Nepal (mofa.gov.np) updated their webpage in 2024, two years after our first scraping in 2022.

As we collect raw HTML content, we also extracted information from `img` tags to identify the pictures used in the articles and downloaded them for image analysis. Since websites use the same `img` tag for logos, icons, etc. we collected distinct paths for images for each website and repeating entries were assumed as design components and excluded. The remaining unique images downloaded and stored for analysis. However, there are also some countries like the United Arab Emirates which provide images loaded with Javascript in the website. In such cases, we were not able to collect and estimate image information from those websites without manual intervention, so we excluded them from the sample for images.

Temporal analysis of diplomatic news

Each government demonstrates varying levels of transparency in disclosing diplomatic information. We have illustrated different characteristics of our textual data in Figure 2(a), which demonstrates the availability of news articles and their concentration over the years. In the figure, we present data from 65 countries that the Lowy Institute includes in its Global Diplomacy Index, except for the European Union [68]. This information is useful for several reasons. First, it signifies how countries structure and systematize their reporting activities based on their current webpage content. Figure 2(a) shows that several countries -such as Russia, the United States, Japan, South Africa, and the United Kingdom- have a more consistent pattern of information sharing. In cases where data coverage is limited, the Internet Archive can be used to complement information. Accordingly, we collected news from the Internet Archive for countries including Iran, the Netherlands, Denmark, and China.

Each government webpage has a specific language option. In Figure 2(b), Spanish, Catalan, Russian, and French -as the main webpage language- are the most prevalent non-English languages in the GlobalDiplomacyNET. Notably, Russian news does not originate from the Russian Federation itself but rather from former Soviet republics such as Kyrgyzstan and partly Azerbaijan. Also, some countries in South and Central America show a strong tendency to publish Spanish news, which indicates that colonial ties can shape diplomatic communication regarding diplomatic news. This shows the impact of cultural proximity among these countries, which is particularly interesting for diplomatic communication. Figure 2(c) illustrates the initial year of publication of diplomatic communications (news). This reveals both the strengths and temporal limitations inherent in our dataset. A significant portion of news articles began in 2014, which still effectively covers the past 11 years. Figure 2(d) demonstrates the verbosity in the dataset, measured by the number of characters in each observation. Some countries, such as Cuba and Colombia, are particularly verbose compared to others. Countries with higher textual verbosity provide more input for the study of inter-state relations. This verbosity is significant also for images to gather more information from the content. Figure 2(e) provides information on visual communication strategies in international politics. The steep power-law distribution reveals that while most diplomatic communications use minimal visual content, a small subset incorporates numerous images.

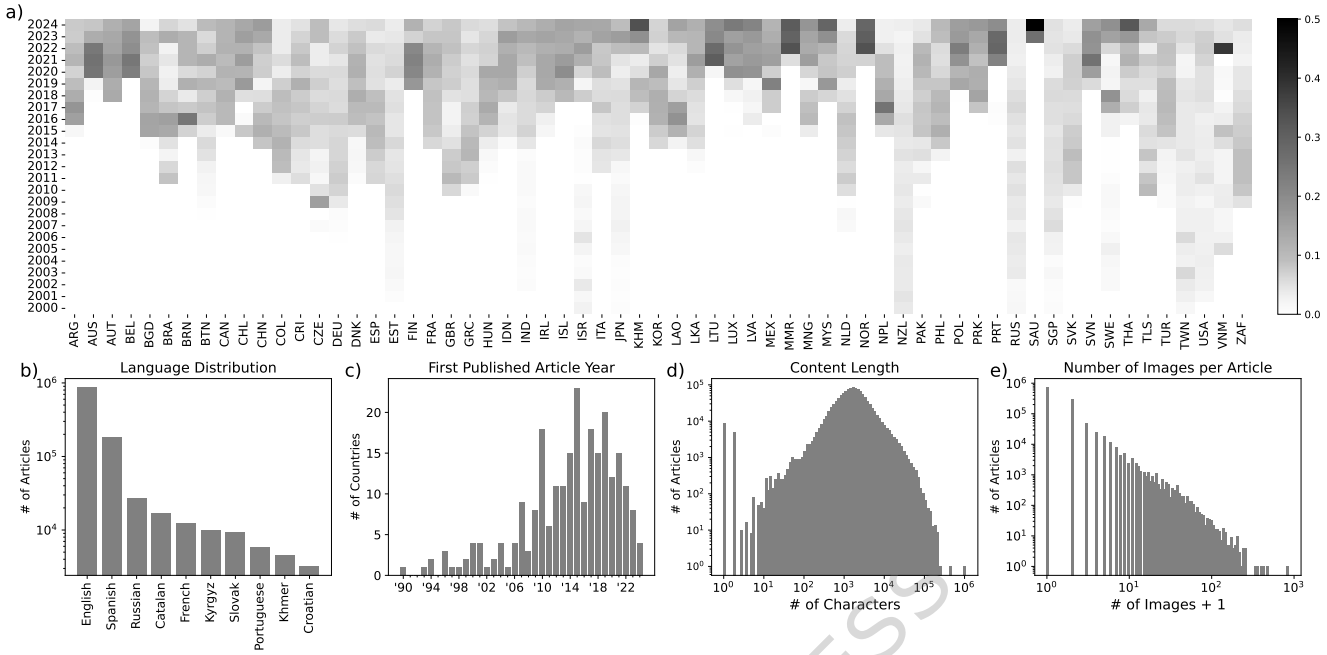


Figure 2. Dataset statistics. Collection of over 1.15 million articles summarized in terms of their temporal spans and content. We report temporal coverage for exemplar countries (a) and the first published articles for all countries collected for the project (c). Most of these reports written in English, followed by Spanish and Russian (b). Reported news can have just a title or as long as 10,000 characters (d) and they can contain images to support them (e).

Detecting entities from news content

Named-entity recognition (NER) procedures are essential for text analysis, since they enable us to measure countries' attention to specific regions, multilateral organizations, and persons. For our systematic analysis, we evaluated diverse family of models ranging from machine learning models trained specifically for this task or general use large language models (LLM) with few-shot prompting for NER task. We focused on three entity types: `PERSON`, `COUNTRY`, and `ORGANIZATION`.

Although most of the news was written in English, some news is collected in other languages. For language detection, we concatenated the title and article body and used the `FastText` model developed by Facebook AI Research, which supports the identification of 176 different languages[69,70]. This model also provides confidence level for detected languages and we only take into account when confidence level is above 0.5. When the detected language for the text is not English, we used Google's Translate using `googletrans` package[71]. News that are longer than 2,000 characters split into smaller chunks and translated texts concatenated later. We preferred Google's Translation API, since we also noticed some countries already integrated their JavaScript widget for adding different languages to their websites.

We evaluated the performance of several NLP libraries and models including NuExtract, Spacy and Gliner and LLMs with different parameter sizes including DeepSeek-R1, Llama, and Mistral. We used Ollama to facilitate local execution of these LLMs. These are among the most advanced LLMs and NLP libraries that demonstrate high performance for NER procedures[72]. Nine researchers from GlobalDiplomacyNET annotated 200 distinct diplomatic news items before the main execution process. These annotations were randomly assigned for evaluator assessment and later used to evaluate model performance. Additionally, each researcher annotated 20 extra news articles with 378 unique entities that were shared among all annotators to ensure consistency in the annotation process. Since annotators can detect entities by providing range of tokens, in some cases additional letters or words can be selected some annotators while others exclude those. Fuzzy string matching used to consider those entities as equivalent. Evaluation of inter-annotator agreement for span prediction tasks like NER is a complex task that needs to address challenges like chance agreement and class imbalance due to un-annotated tokens. In literature different measures like Fleiss' kappa, Krippendorff's alpha, and F1 scores for pairwise comparisons were proposed [73–76]. We use different measures to assess inter-annotator agreement. A straightforward analysis for Fleiss' kappa offers score of 0.77, which demonstrates a substantial level of inter-annotator agreement particularly for political texts [73]. We also compute Krippendorff's alpha in two settings since this measure can also take into account missing annotations: i) defining a separate label to penalize coder for missing an entity and ii) considering missing data for a coder if they are unable to detect an entity and not offering any label. Both cases lead to substantial agreement: 0.98 for recommended usage where we treat non-labeled tokens as missing

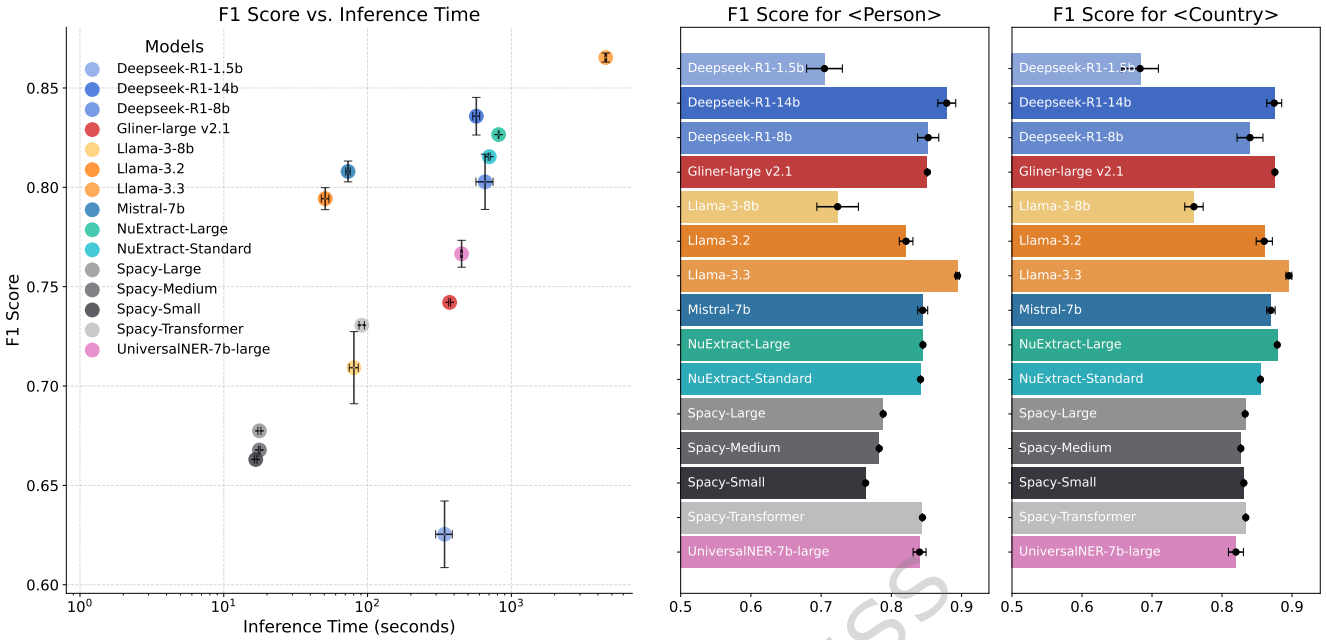


Figure 3. Model comparison for named-entity detection. Models were compared in terms of their inference time for 200 random news and detection performance with 10 repetitions to calculate error bars (a). Detection performance for `PERSON` and `COUNTRY` reported with models' corresponding F1 scores (b,c).

annotation and 0.69 for more conservative estimate of treating missing entities as different labels. Analysis by considering the annotations of one annotator as the ground-truth and annotations of another as the predictions is used to calculate F1 scores [74,75]. For pairwise evaluation of annotations, we obtained F1 score of 0.78 ± 0.057. By incorporating different measures of agreement, we can suggest that manual annotations created to evaluate different NER tools are in a reasonable quality.

The NER tool that is closest to our annotation is analyzed in Figure 3. To quantify detection performances, we used F1 score measure. Since these models used to extract entities over a million documents, models execution time is also relevant. `Llama-3.3` offers the best performance, but its inference time is almost an order of magnitude higher than other models. Because of this limitation, we selected `DeepSeek-R1-14b`[77] for its efficient inference time and performance of 0.836 F1 score. These findings also corroborate LLMs performances measured by other scholars in different research areas [78].

Since detected entities across documents can vary (e.g. Recep Tayyip Erdoğan, Tayyip Erdoğan, and President Erdoğan are all the same person), we use Wikipedia projects to resolve disambiguities and linking with a unique Wikidata QID (Wikidata unique identifier) that we will refer to as Wikidata QID. For each unique person entity, we collected responses from Wikipedia Search API [79] and correspondingly Wikidata Search API and built a bi-partite network of entity strings and resolved Wikidata QIDs. We calculate the mapping between person entity strings and unique Wikipedia pages based on fuzzy-string matching between entity and Wikipedia page title as well as frequency of referral to Wikipedia pages. We later convert the Wikipedia URIs obtained from fuzzy matching to Wikidata QIDs. This way we assigned a Wikidata QID to all person entity strings that appear at least 10 times in our corpus. As for the countries and organizations we used Wikidata Search API directly for matching the entity string to page title. We used Wikidata's structured knowledge base to compile the set of international organizations we wanted to include. After both country and organization entities were resolved, we validated the results manually as the entity space was quite small.

Image Analysis

Advances in deep learning make systematic analysis possible to study large-volume of image data. To extract information from the images, we consider task-specific models for detecting people in the images and inferring their human characteristics such as gender. The Vision Language Models also offer opportunities to interact with the content through prompting with natural language. These models are computationally more expensive in terms of time and computational resources, but they are quite capable of answering simple image understanding questions.

The analysis focused on extracting information on gender and number of people shared on diplomatic news websites. The images were downloaded using the source URLs identified for unique images within the HTML files. For processing, we use pretrained models to extract information like `YOLOv8` large model[80] for detecting humans in the images. These human images

a) news.jsonl


```
{
  "id": "bacfe48df5be372e9ff880785e5e4bcf",
  "url": "https://www.tccb.gov.tr/en/news/542/91796/turkey-has-th...",
  "date": "2018-03-16",
  "title": "“Turkey has the power to take action for its own secu...",
  "content": " Speaking at the AK Party’s provincial congress in ...",
  "lang": "en",
  "title_original": "",
  "content_original": "",
  "entities": {
    "persons": ["Recep Tayyip Erdoğan","Mr. Obama","President Er..."],
    "countries": ["United States","U.S.","Iraq","Turkey","Syria"],
    "organizations": ["United Nations"]
  },
  "wikidata_qids": {
    "persons": ["Q39259","Q76","Q39259"],
    "countries": ["Q30","Q30","Q796","Q43","Q858"],
    "organizations": ["Q1065"]
  }
}
```


b) images.jsonl


```
{
  "id": "d963463741063cd5a4899172fc689fba",
  "news-id": "bacfe48df5be372e9ff880785e5e4bcf",
  "url": "https://www.tccb.gov.tr/ImageResizer/CropImage?w=-1&h=-...",
  "male-count": 6,
  "female-count": 2
}
```


Figure 4. Samples from GlobalDiplomacyNET data records. Key-value pairs from the JSON files that contain records for news (a) and images (b).

were later analyzed by a pre-trained gender classification model on HuggingFace [81].

The exploratory analysis investigated two primary visual characteristics: overall gender composition and the number of people included in an image. This enabled us to measure and identify trends and disparities among countries in the portrayal of gender on diplomatic websites. In GlobalDiplomacyNETdataset, we provide statistics about the images and link those images with the articles shared in the dataset. We hope that the future research will be conducted with vision models and information extracted for the GlobalDiplomacyNETdataset. The potential to combine multiple datasets and investigate visual data is a promising research direction.

Data Records

Dataset offered by the GlobalDiplomacyNET project is available on Harvard Dataverse[82]. Aggregated data from 156 countries and 248 websites include 1,163,905 news documents and 1,187,152 images. In our dataset, we followed a specific naming convention to define folders and all relevant data places in that directory. Folder names composed of three components: `<COUNTRY-CODE>_<TYPE>_<COUNT>`. Country names are converted to standard alpha-3 country codes. We also distinguish websites for Executives (exec) and Ministries of Foreign Affairs (mofa) by `<TYPE>`. Since some countries may have separate websites for different presidents or time intervals, we also added `<COUNT>` key, which is only available when there are more than one version for that website. This way, users of our dataset can analyze particular countries of interest separately. In Figure 4, we present samples records from the GlobalDiplomacyNET dataset. Within each country folder, we provide the following files:

`news.jsonl`: Each line in this file contains a separate news as JSON object. These objects contain a unique identifier (`id`), news details (`url`, `date`, `title`, and `content`). Since some countries publish articles in languages other than English, we add a key for the detected language (`lang`) and original content accessible with `content-original` and translated content is placed in the `content`. We also share detected entities (`entities`) and their corresponding unique Wikidata QIDs (`wikidata-qid`).

`images.jsonl`: For computational image analysis, we provide records that contains a unique identifier for an image (`id`), ID from `news.jsonl` to link images and news (`news-id`), and source of the image (`url`) as main information. We also provide results of the image analysis, detected number of males (`male-count`) and females (`female-count`) in the picture.

On the same directory as the country folders we also provide a summarizing file:

`summary_statistics.xlsx`: The summary statistics table lists all URLs associated with each website -as our sources- in the Harvard Dataverse dataset. We report the name of the country, its 3-letter ISO Alpha code, type of the website (“exec” or “mofa”), number of news, time-span of the news, fraction of non-English content, number of images, and female ratio.

Technical Validation

Validating named-entities within the news content temporally

Diplomatic texts mention different countries, regions, international organizations, politicians, and diplomats. Systematically identifying these entities is useful to capture governments’ geographical, individual, and multilateral attention. For instance,

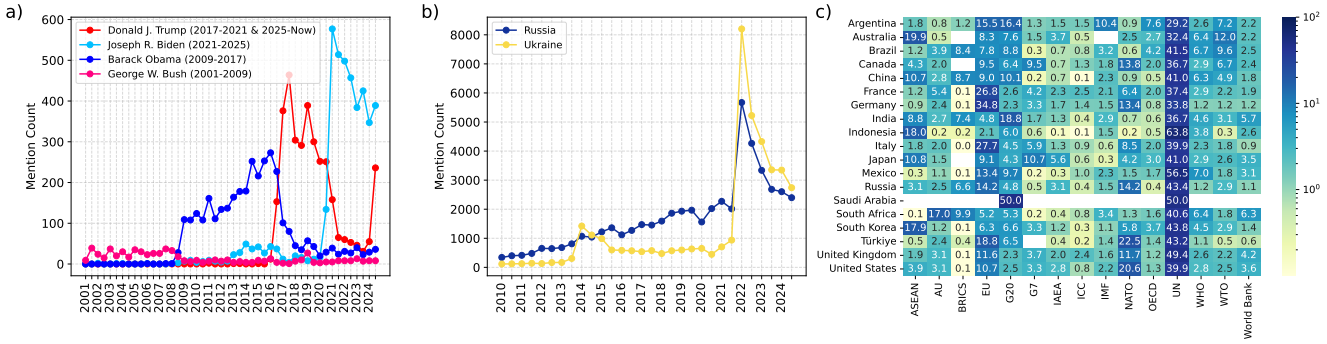


Figure 5. Global attention for named-entities over time. Entities extracted for `PERSON` (a), `COUNTRY` (b), and `ORGANIZATION` (c). We demonstrate number of mentions for the last four US presidents (a), Ukraine and Russia (b), and occurrences of Major International Organizations in the news of G20 Countries (c).

we can study co-occurrences of these countries or individuals to quantify the relations between these entities and it is also possible to filter these networks with temporal slides to investigate changes over time. This data can also test which multilateral institutions are becoming more salient over time. For instance, countries' shifting focus toward or away from BRICS serves as an indicator of their attention to such organizations, showing possible geoeconomic adjustments. This is particularly timely and relevant given the growing regionalization in international politics[83].

Figure 5 illustrates how often various named entities appear in our collection for interesting exemplary cases. Figure 5(a) tracks mention frequencies for the last four U.S. Presidents -Joseph R. Biden, Donald J.Trump, George W. Bush, and Barack Obama- revealing that each has distinct spikes in coverage corresponding to their presidential terms. It is significant that countries (other than the US) mentioned Obama even after his presidential tenure, which is not the case for George W.Bush. Donald Trump's first presidential term attracted a sharp surge in diplomatic attention. However, after his term ended, this attention decreased sharply. Figure 5(b) compares Russia (in blue) and Ukraine (in yellow) mentions, showing significant increases at key geopolitical moments (e.g., Ukraine's sharp rise around 2014–2015 and again in 2022). Diplomatic news often contains early signals that precede actual formal conflicts. For instance, a surge in the monthly or daily mentions of Ukraine ahead of 2022 might reflect an escalation in rhetorical salience long before an actual invasion is triggered. Figure 5(b) also shows that mentions of Ukraine and Russia by other countries started to increase before the invasion. Researchers can adjust country-attention according to their own research in different regional contexts such as Taiwan, West Bank and Gaza, and others.

Finally, Figure 5(c) provides a heatmap of major international organizations (columns) as referenced by different G20 countries (rows), where cell colors indicate the intensity of mentions (ranging from lower mentions in lighter shades to higher mentions in darker shades). The United Nations (UN), NATO, the European Union (EU), and The Association of Southeast Asian Nations(ASEAN) emerge as the most salient international organizations in the diplomatic communications of G20 countries. A high frequency of the UN mentions are expected, and Mexico, United Kingdom, Indonesia, South Korea emerge as the countries that mention the UN most. A prominent example is Russia, Turkiye, United States and Canada, and Germany, which mention NATO at a higher rate than other countries. Indonesia and India are the countries that refer to NATO the least in their diplomatic communications. The ASEAN is also mostly mentioned by Asian countries. Moreover, as a non-ASEAN partner country, China's diplomatic communications demonstrate remarkable focus on this regional organization. In China's diplomatic news coverage, ASEAN appears alongside other significant international bodies such as BRICS, G20, the United Nations, and the European Union, which can show China's strategic prioritization of Southeast Asian regional engagement. It is also significant to mention that the WTO is mostly referenced by Australia, Argentina, Brazil and Mexico, which might demonstrate middle powers' export oriented concerns. These analysis offer meaningful interpretation of temporal patterns of named-entities identified within diplomatic texts for domain experts. Information extracted from GlobalDiplomacyNET can be used for future research, and researchers can also utilize full text provided with the dataset for more in-depth analysis.

Computational Image Analysis

In addition to textual verbosity, we also analyze governments' practices of sharing images from diplomatic events and press releases. Researchers can pose several questions to study the information conveyed through image content. To minimize potential bias arising from insufficient data in this validation study, we excluded countries with fewer than 100 images. This threshold was set based on the distribution of images available per country, which ensures the statistical reliability and representativeness of the results. It is important to underline that we only collect images if the news HTML contains an `<img>`

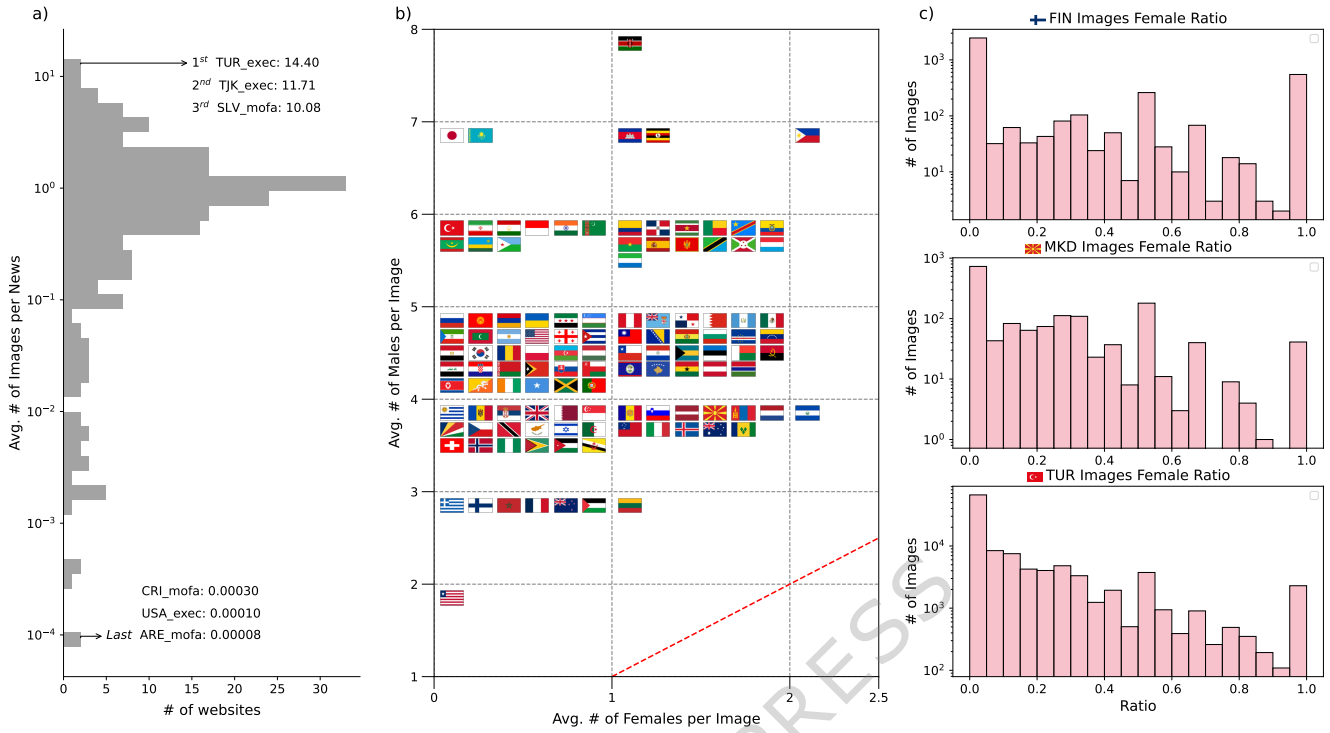


Figure 6. Analysis of images and participation in diplomatic events based on gender. 1.18 million images summarized in terms of their frequency per news (a). Average number of males and females across countries (b). Selected countries' image distribution and gender fraction (c).

tag. Images embedded in JavaScript are therefore not captured.

Based on Figure 6(a), we find that countries such as Türkiye, Tajikistan, and El Salvador have the highest number of images per news. These statistics are an aggregation that originates from both the ministry of foreign affairs and chief executive websites. We generally detected that chief executive webpages tend to have a higher number of images per news article when compared to the ministries of foreign affairs. We also want to point out that some of these countries, such as Türkiye, chief executive webpages publish information on the domestic presidential agenda. In such cases, diplomatic news is blended with domestic issues under certain circumstances. Indeed, some of these pictures come from public meetings and government propaganda, which can substantially inflate the number of images. Therefore, such circumstances do not directly imply the "diplomatic" side of image sharing. However, this news content is in English and directly communicates with the global audience. We can interpret this finding as part of public diplomacy in which countries signal their political resilience and public support. Moreover, if researchers desire to acquire visuals (or texts) related to just foreign politics from chief executive webpages, they can use GlobalDiplomacyNET's country-entities to filter foreign-related visuals and news content.

Figure 6(b) shows a significant trend in gender-dimension of image-sharing. The plot shows countries relative to the diagonal line, which represents absolute gender-equality.. The figure shows that most flags lie above the diagonal, which indicates that in the vast majority of these news articles, men appear more often than women. Countries are concentrated between 4 and 5 men per image. We can suggest that the presence of women in diplomatic and political visuals is always less than men's in all countries in our dataset. El Salvador and Lithuania are the closest countries to the equality line. European and Latin American countries seem to include more women in their diplomatic news as seen in Figure 6(b). Researchers can employ our image analysis across multiple time periods and geographical contexts by appealing different variables. For instance, the democracy and gender-dimension of diplomacy have been significant issues in the literature, where GlobalDiplomacyNET image analysis can leverage such domain-specific research fields. Our current study's scope is limited to 2024 data, but future work could reveal valuable longitudinal trends and cross-cultural patterns on countries' image-sharing habits. Figure 6(c) reveals the distribution of the female ratio for three examples including Finland, North Macedonia and Turkey on a log y-scale. In all three countries, images with many men are far more common than images with many women. Finland is relatively more balanced but still male-skewed. Türkiye has the longest right tail for men, because the crowd events reaching more than 30 men per image on average. Moreover, women presentation generally decreases if images include a crowded population.

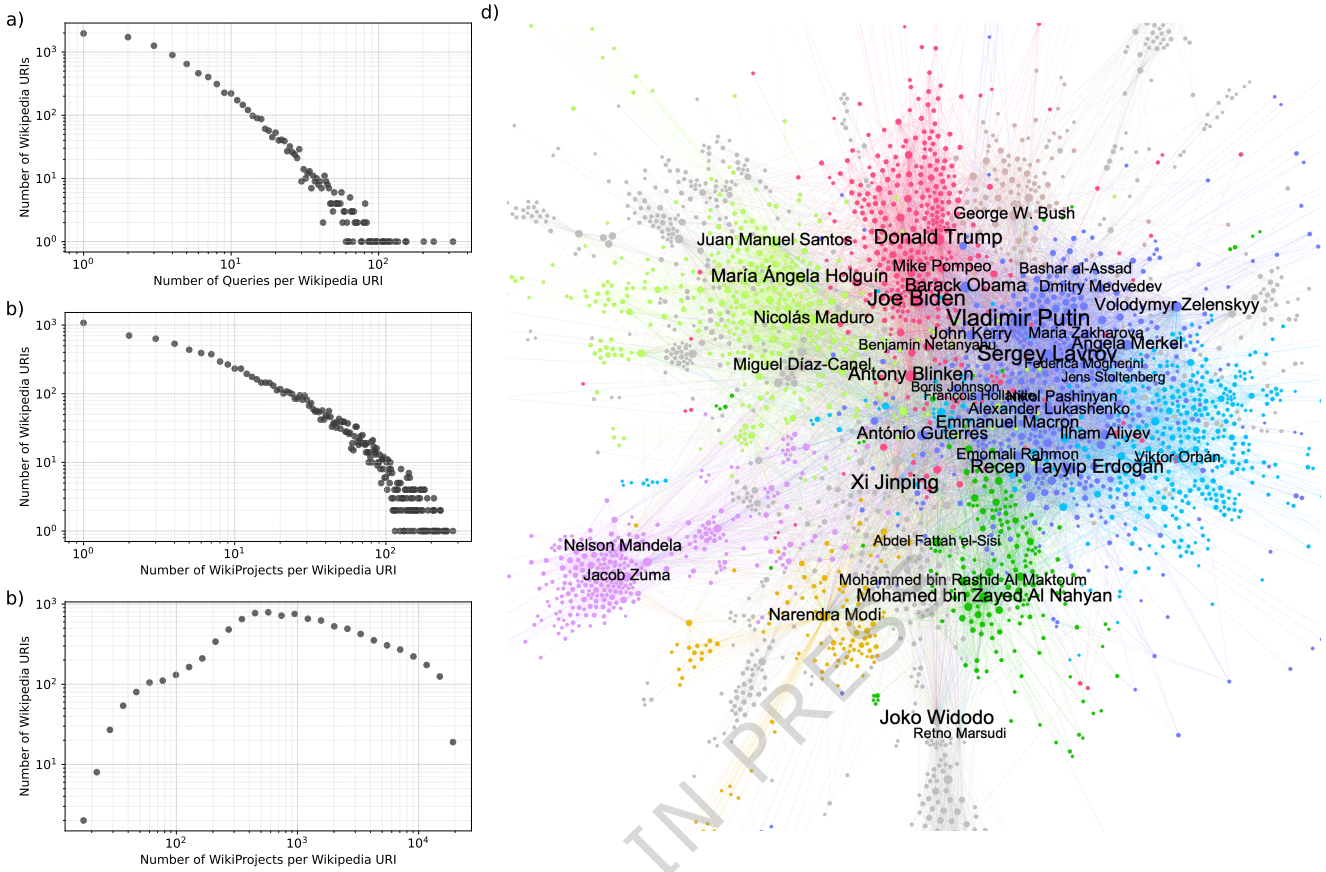


Figure 7. Network analysis of individuals through Wikipedia URIs. Same Wikipedia IDs can be referred with different ways in the news and their distribution follows a power-law (a). To assess the popularity of each person, we collected Wikimedia project for different languages (b) and also calculate the content length distribution for English page (c). The co-occurrence network presents people in the nodes and their interaction frequency represented as edge weight (d).

Validating network of people from entity co-occurrences

Network science offers tools to analyze interactions between different entities, here we investigate co-occurrences of people with the same news. This network representation points clusters of individuals that reflect mostly the geographical associations, but also their interactions due to diplomatic agenda and shared organizational memberships. One can also study such network overtime to investigate when certain world leaders or foreign ministers begin to appear together in diplomatic communications. Such analysis may signal the formation or strengthening of bilateral ties. Conversely, an interruption in mentions of a particular individual following a policy dispute can forecast increased tensions.

In Figure 7, we present results on co-occurrence network of individual people appeared on diplomatic news. Nodes in this network correspond to different Wikipedia URIs and links between them capture the frequency of appearing on the same news. Use of the Wikipedia API helped us disambiguate different ways to express same political figures in the text and map them into single entities. We plot the distribution for the number of unique entities mapped to particular Wikipedia URI in Figure 7(a). Most political figures mentioned in a few different ways; however, some individuals were mentioned more than 100 times. When we inspect those individuals, we noticed they take multiple titles over the year or they adopt using multiple family names to show their ancestry. For instance, President of the United Arab Emirates Mohamed bin Zayed Al Nahyan is most frequently referred to as “his highness sheikh mohamed bin zayed al nahyan”, “h.h. sheikh hamed bin zayed al nahyan” or “sheikh mohammed bin zayed al nahyan”.

It is also known that not all political figures are equally represented. Some of them are more prominent than others, and we can quantify that by analyzing how many different Wikimedia projects they have a page (see Figure 7(b)) and what the length of the content is in the English edition of Wikipedia as shown in Figure7(c). When we inspect the top leaders based on the different Wikimedia project they covered, US presidents like Donald Trump, Barack Obama, and Joe Biden were among top-10. We also detected founding leaders like Nelson Mandela and Mustafa Kemal Atatürk in the list. One can collect further

information about those figures and only analyze the politically active individuals.

In Figure 7(d), we visualize the largest connected component of the co-occurrence network. We highlight the nodes corresponding to politicians who are among the top 30 for either having the highest degree centrality or node strength. This network encompasses all years in our dataset and provides insights into international political trends. Clusters correspond in part to regional groupings. For instance, Latin America's political figures are concentrated in the green nodes, such as Juan Manuel Santos, Angela Holguin, Nicolas Maduro, and Miguel Diaz-Canel. This co-occurrence can result from these leaders' meetings, political interactions and mentions in the news. However, diplomatic news includes cables and many other forms of person-mentions for political figures. António Guterres emerges as a central political figure in the network, which makes sense due to his position as the General-Secretary of the United Nations.

A co-occurrence network can also be related to political figures' salience instead of their regional position. Zelenskyy, Medvedev, Bashar al-Assad, Putin, Angela Merkel, and Lavrov mostly co-occur together in countries' foreign news. In this regard, Al-Assad -Syria's overthrown leader- does not emerge as part of the Middle-Eastern cluster in the network. In other words, the mentioned characters' deviation from their regional cluster seems to parallel their corresponding salience in global affairs. Within the United States' high-level representatives, Anthony Blinken is closest to the epicenter of the co-occurrence network, meaning that his name co-occurs with other countries' political figures more than other American politicians. It is also interesting that China's president, Xi Jinping, is very close to the center of the network, implying China's global diplomatic hinterland. The network analysis offers tools for validating how different entities interacts and richness of the data offered in GlobalDiplomacyNET. Our dataset can be used to study temporal dynamics between political figures or in combination with other data sources about political leaders and regimes to investigate novel research questions.

Data Availability

We released the entire GlobalDiplomacyNET dataset on Harvard Dataverse (doi.org/10.7910/DVN/HYJDE0). Users can download the full text and additional content that we offer from diplomatic news such as entities, image features for different countries websites. All data are public records and were scraped from publicly accessible webpages of ministries of foreign affairs and chief executive offices, whose URLs are listed in the Summary Statistics table. All source webpages are publicly available without access restriction. No personal data were collected, and our usage and data mining practices strictly comply with the open science principles, re-use provisions under EU Directives 2019/1024[66], 2019/790[67], and other recommendations from OECD Member Countries, including the UK, and commonwealth countries (Canada, Australia).

Usage Notes

We also developed an interactive website for our users to inspect different dimensions of our analysis and access our dataset. Our GlobalDiplomacyNET website is online at globaldiplomacy.net. We plan to keep the dataset updated by executing the data collection pipelines yearly. Since websites can change their design, we plan to adjust those changes in our code base. One of the known limitations of diplomatic websites is their limited capacity to archive their content. Most websites delete older posts for no transparent reason or miss them in website updates. Wayback Machine of the Internet Web Archive could be an alternative; however, most of the sites only have snapshots of their landing pages and links to diplomatic news were not captured by the crawlers.

Code availability

We provide a GitHub repository to share codes for the analysis conducted in this paper: github.com/ViralLab/GlobalDiplomacyNet-Dataset. For webscraping we developed custom codes for each website. Selenium and requests libraries were used to retrieve data and BeautifulSoup4 package used for parsing HTML content.

Acknowledgments

We thank members of VRL Lab at Sabanci University for their valuable feedback. Melis Gemalmaz, Yunus Tan Kerestecioğlu, and Yağız Demirbaş for their contribution to webscraping. This work was supported by TUBITAK under the grant agreement 223K173. We also thank TUBITAK 121C220 for their partial support.

Author contributions statement

B.B., A.D., A.T.K, K.G.S., D.T., F.G.Y, N.M., O.V., and Y.Z. contributed to web scraping. B.B., A.D., A.T.K, K.G.S., F.G.Y, M.H, N.M.and D.T annotated sample data. N.M curated the list of institutions and country sample. K.G.S. was responsible for data management and reproducibility. F.G.Y. and M.H. experimented with NER models. M.H. implemented NER pipelines.

F.G.Y. implemented translation pipelines. B.B., A.T.K, and K.G.S. contributed to image analysis. F.G.Y. and O.V. conducted network analysis. A.T.K developed the interactive website. N.M. and O.V. developed the concept and the methodologies. Manuscript is written by N.M and O.V. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests

References

1. Grimmer, J. & Stewart, B. M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Polit. Analysis* **21**(3), 267–297. doi:10.1093/pan/mps028 (2013).
2. Laver, M., Benoit, K. & Garry, J. Extracting Policy Positions from Political Texts Using Words as Data. *Am. Polit. Sci. Rev.* **97**(2), 311–331. doi:10.1017/S0003055403000698 (2003).
3. Nguyen, D., Liakata, M., DeDeo, S., Eisenstein, J., Mimno, D., Tromble, R., Summers-Stay, D., & Vilares, D. How we do things with words: Analyzing text as social and cultural data. *Front. Artif. Intell.* **3**, 62. doi:10.3389/frai.2020.00062 (2020).
4. Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., & Ungar, L. H. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE* **8**(9), e73791. doi:10.1371/journal.pone.0073791 (2013).
5. Young, L. & Soroka, S. Affective News: The Automated Coding of Sentiment in Political Texts. *Polit. Commun.* **29**(2), 205–231. doi:10.1080/10584609.2012.671234 (2012).
6. Wilkerson, J. & Casas, A. Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. *Annu. Rev. Polit. Sci.* **20**, 529–544. doi:10.1146/annurev-polisci-052615-025542 (2017).
7. Baturo, A., Dasandi, N. & Mikhaylov, S. J. Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Res. & Polit.* **4**(2), 2053168017712821. doi:10.1177/2053168017712821 (2017).
8. Watanabe, K. & Zhou, Y. Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. *Soc. Sci. Comput. Rev.* **40**(2), 346–366. doi:10.1177/0894439320907027 (2022).
9. Dahlgren, P. M.. mediacommtools: Tools/resources for journalism, media, communication, computational social science. *GitHub repository* (2022). https://github.com/peterdalle/mediacommtools. Accessed November 13, 2025.
10. Ünver, H. A. Computational international relations: What can programming, coding and internet research do for the discipline? *All Azimuth: A J. Foreign Policy Peace* **8**(2), 157–182. doi:10.20991/allazimuth.476433 (2019).
11. Manor, I. Exploring the semiotics of public diplomacy. *CPD Perspectives* **2** (2022).
12. Joo, J. & Steinert-Threlkeld, Z. C. Image as Data: Automated Visual Content Analysis for Political Science. *arXiv preprint arXiv:1810.01544*. doi:10.48550/arXiv.1810.01544 (2018).
13. Leetaru, K. & Schrodt, P. A. GDELT: Global Data on Events, Location and Tone, 1979–2012. In *ISA Annual Convention*, vol. 2, 1–49 (The GDELT Project, 2013). https://www.gdeltproject.org/.
14. Carson, A., Min, E. & Van Nuys, M. Racial tropes in the foreign policy bureaucracy: A computational text analysis. *Int. Organ.* **78**, 189–223. doi:10.1017/S0020818324000146 (2024).
15. Yonamine, J E. Predicting future levels of violence in Afghanistan districts using GDELT. *Unpubl. manuscript* (2013).
16. Hong, L. & Davison, B. D. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, 80–88. doi:10.1145/1964858.1964870 (2010).
17. Tonidandel, S., Summerville, K. M., Gentry, W. A. & Young, S. F. Using structural topic modeling to gain insight into challenges faced by leaders. *The Leadersh. Q.* **33**(5), 101576. doi:10.1016/j.leaqua.2021.101576 (2022).
18. Erhard, L., Hanke, S., Remer, U., Falenska, A. & Heiberger, R. H. PopBERT: Detecting Populism and Its Host Ideologies in the German Bundestag. *Polit. Analysis* **33**(1), 1–17. doi:10.1017/pan.2024.12 (2025).
19. Dada, S., Ashworth, H. C., Bewa, M. J. & Dhatt, R. Words matter: political and gender analysis of speeches made by heads of government during the COVID-19 pandemic. *BMJ Glob. Heal.* **6**, e003910. doi:10.1136/bmjgh-2020-003910 (2021).
20. Curiskis, S. A., Drake, B., Osborn, T. R. & Kennedy, P. J. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Inf. Process. & Manag.* **57**, 102034. doi:10.1016/j.ipm.2019.04.002 (2020).

21. Fisher, S., Klein, G. R. & Codjo, J. FOCUSDATA: Foreign policy through language and sentiment. *Foreign Policy Analysis* **18**, orac002. doi:10.1093/fpa/orac002 (2022).

22. Connelly, M. J., Hicks, R., Jervis, R., Spirling, A. & Suong, C. H. Diplomatic documents data for international relations: The Freedom of Information Archive Database. *Confl. Manag. Peace Sci.* **38**(6), 762–781. doi:10.1177/0738894220930326 (2021).

23. Katagiri, A. & Min, E. The credibility of public and private signals: A document-based approach. *Am. Polit. Sci. Rev.* **113**, 156–172. doi:10.1017/S0003055418000643 (2019).

24. Baykaldi, S. Examining public diplomacy message strategies during times of crisis communication: An analysis of recent major Turkey–Russia crises (Michigan State University, 2023).

25. Casler, D., Connelly, M. & Hicks, R. Trading with Frenemies: How Economic Diplomacy Affects Exports. *Int. Stud. Q.* **68**, sqae098. doi:10.1093/isq/sqae098 (2024).

26. Demir, Y., Keskin, İ. & Kutluoğlu, M. H. Diplomatic archives: A theoretical evaluation. In *Current Issues in Archival Science*, pp. 26–43. doi:10.26650/B/SSc19SSc21.2025.003.002 (Istanbul University Press, 2025).

27. Romero, V., Toselli, A. H., Vidal, E., Sánchez, J. A., Alonso, C. & Marqués, L. Modern vs Diplomatic Transcripts for Historical Handwritten Text Recognition. In *New Trends in Image Analysis and Processing – ICIAP 2019* (eds. Cristani, M., Prati, A., Lanz, O., Messelodi, S., Sebe, N.). **Lecture Notes in Computer Science**, vol 11808, pp. 103–114, Springer, Cham. doi:10.1007/978-3-030-30754-7$_1$1 (2019).

28. Allen, D. & Connelly, M. Diplomatic history after the Big Bang: Using computational methods to explore the infinite archive. In *Explaining the History of American Foreign Relations*, pp. 74–101, Cambridge University Press, Cambridge, 3rd ed. doi:10.1017/CBO9781107286207.006 (2016).

29. Heseltine, M. & Clemm von Hohenberg, B. Large language models as a substitute for human experts in annotating political text. *Res. & Polit.* **11**, 20531680241236239. doi:10.1177/20531680241236239 (2024).

30. Lebovic, J. H. & Saunders, E. N. The Diplomatic Core: The Determinants of High-Level US Diplomatic Visits, 1946-2010. *Int. Stud. Q.* **60**(1), 107–123. doi:10.1093/isq/sqv008 (2016).

31. Bayer, R. Diplomatic exchange dataset, 1817–2005 (v2006.1). *Correl. War* (2006).

32. Davis, C. L. International institutions and issue linkage: Building support for agricultural trade liberalization. *Am. Polit. Sci. Rev.* **98**, 153–169. doi:10.1017/S0003055404001066 (2004).

33. Balci, A. & Pulat, A. Love, money, or fame? Determinants of Turkey's leader visits. *Int. Stud. Q.* **68**, sqad104. doi:10.1093/isq/sqad104 (2024).

34. Malis, M. & Smith, A. State Visits and Leader Survival. *Am. J. Polit. Sci.* **65**(1), 241–256. doi:10.1111/ajps.12520 (2021).

35. Moyer, J. D., Meisel, C. J., Szymanski-Burgos, A., Scott, A. C., Casiraghi, M. CM., Kurkul, A., Hughes, M., Kettlun, W., McKee, K.X. & Matthews, A.S. When Heads of Government and State (HOGS) Fly: Introducing the Country and Organizational Leader Travel (COLT) Dataset Measuring Foreign Travel by HOGS. *Int. Stud. Q.* **69**(2), sqaf013. doi:10.1093/isq/sqaf013 (2025).

36. Maoz, Z., Johnson, P. L., Kaplan, J., Ogunkoya, F. & Shreve, A. P. The Dyadic Militarized Interstate Disputes (MIDs) Dataset Version 3.0: Logic, Characteristics, and Comparisons to Alternative Datasets. *J. Confl. Resolut.* **63**(3), 811–835. doi:10.1177/0022002718784158 (2019).

37. Moyer, J. D., Turner, S. D. & Meisel, C. J. What are the drivers of diplomacy? Introducing and testing new annual dyadic data measuring diplomatic exchange. *J. Peace Res.* **58**(6), 1300–1310. doi:10.1177/0022343320929740 (2021).

38. Moyer, J D, Bohl, D & Turner, S D. DIPLOMETRICS: Diplomatic Representation Data Codebook (2022).

39. Rhamey, P., Cline, K., Thorne, N., Cramer, J., Miller, J. L. & Volgy, T. J.. The Diplomatic Contacts Database. *Tucson: Sch. Gov. Public Policy, Univ. Ariz.* (2013). Version 3.0.

40. Diplomatic Pulse. https://diplomaticpulse.org/ (2025).

41. Colladon, A. F. & Vestrelli, R. A Python tool for reconstructing full news text from GDELT. *arXiv preprint arXiv:2504.16063*. doi:10.48550/arXiv.2504.16063 (2025).

42. Goodman, S., Zhang, S., Malik, A. A., Parks, B. C. & Hall, J. AidData's Geospatial Global Chinese Development Finance Dataset. *Sci. Data* **11**, 529. doi:10.1038/s41597-024-03341-w (2024).

43. Scholz, S., Weidmann, N. B., Steinert-Threlkeld, Z. C., Keremoğlu, E. & Goldlücke, B. Improving computer vision interpretability: Transparent two-level classification for complex scenes. *Polit. Analysis* **33**, 107–121. doi:10.1017/pan.2024.18 (2024).

44. Wang, Y., Li, Y. & Luo, J. Deciphering the 2016 US presidential campaign in the Twitter sphere: A comparison of the Trumpists and Clintonists. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, 723–726. doi:10.1609/icwsm.v10i1.14783 (2016).

45. Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B. & Hsiang, S. A generalizable and accessible approach to machine learning with global satellite imagery. *Nat. Commun.* **12**, 4392. doi:10.1038/s41467-021-24638-z (2021).

46. Holmes, M. Digital diplomacy: Projection and retrieval of images and identities. *The Oxf. Handb. Digit. Dipl.* 29 (2024).

47. Hellmann, O. & Oppermann, K. Photographs as instruments of public diplomacy: China's visual storytelling during the Covid-19 pandemic. *The Hague J. Dipl.* **17**(2), 177–215. doi:10.1163/1871191X-bja10097 (2022).

48. Niklasson, B. & Towns, A. E. Diplomatic gender patterns and symbolic status signaling: Introducing the GenDip dataset on gender and diplomatic representation. *Int. Stud. Q.* **67**, sqad089. doi:10.1093/isq/sqad089 (2023).

49. Niner, S., Cummins, D., Sahin, S. B., Mulder, S. & Morrison, E. Women's political participation in post-conflict settings: The case of Timor-Leste. *Asian Stud. Rev.* **46**(2), 293–311. doi:10.1080/10357823.2021.1973365 (2022).

50. Towns, A. & Niklasson, B. Gender, international status, and ambassador appointments. *Foreign Policy Analysis* **13**(3), 521–540. doi:10.1093/fpa/orw039 (2017).

51. Bailey, M. A., Strezhnev, A. & Voeten, E. Estimating dynamic state preferences from United Nations voting data. *J. Confl. Resolut.* **61**(2), 430–456. doi:10.1177/0022002715595700 (2017).

52. Wang, Y. & Stone, R. W. China visits: A dataset of Chinese leaders' foreign visits. *The Rev. Int. Organ.* **18**, 201–225. doi:10.1007/s11558-022-09459-z (2023).

53. Kaufmann, D., Kraay, A. & Mastruzzi, M. The Worldwide Governance Indicators: Methodology and Analytical Issues. *Hague J. on Rule Law* **3**(2), 220–246. doi:10.1017/S1876404511200046 (2011).

54. Marshall, M G & Gurr, T R. Polity5: Political regime characteristics and transitions, 1800–2018. *Cent. for Syst. Peace* **2** (2020).

55. Coppedge, M., Gerring, J., Knutsen, CH., Lindberg, SI., Teorell, J., Altman, D., Angiolillo, F., Bernhard, M., Cornell, A. & Fish, MS. *et al.* V-Dem [Country–Year/Country–Date] Dataset v15. *Var. Democr. (V-Dem) Proj.* doi:10.23696/vdemds25 (2025).

56. Raleigh, C., Linke, A., Hegre, H. & Karlsen, J. Introducing ACLED: An Armed Conflict Location and Event Dataset. *J. Peace Res.* **47**(5), 651–660. doi:10.1177/0022343310378914 (2010).

57. Davies, S., Pettersson, T., Sollenberg, M. & Öberg, M. Organized violence 1989–2024, and the challenges of identifying civilian victims. *J. Peace Res.* **62**(4), 1223-1240. doi:10.1177/00223433251345636 (2025).

58. Strauß, N., Kruikemeier, S., van der Meulen, H. & van Noort, G. Digital diplomacy in GCC countries: Strategic communication of Western embassies on Twitter. *Gov. Inf. Q.* **32**(4), 369–379. doi:10.1016/j.giq.2015.08.001 (2015).

59. Collins, S. D., DeWitt, J. R. & LeFebvre, R. K. Hashtag diplomacy: Twitter as a tool for engaging in public diplomacy and promoting US foreign policy. *Place Branding Public Dipl.* **15**(2), 78–96. doi:10.1057/s41254-019-00119-5 (2019).

60. Danziger, R. & Schreiber, M. Digital diplomacy: Face management in MFA Twitter accounts. *Policy & Internet* **13**(4), 586–605. doi:10.1002/poi3.269 (2021).

61. Zhang, Y., Lukito, J., Suk, J. & McGrady, R. Trump, Twitter, and Truth Social: How Trump used both mainstream and alt-tech social media to drive news media attention. *J. Inf. Technol. & Polit.* **22**, 229–242. doi:10.1080/19331681.2024.2328156 (2025).

62. Bull, H. The Anarchical Society: A Study of Order in World Politics (Bloomsbury Publishing, 2012).

63. Berridge, G R. Diplomacy: Theory and Practice (Springer Nature, 2022).

64. Axworthy, L. The Political Actors: President, Prime Minister, and Minister of Foreign Affairs. In *The Oxford Handbook of Modern Diplomacy* (eds. Cooper, A. F., Heine, J. & Thakur, R.). (Oxford University Press, Oxford, 2013).

65. European Commission Open science. https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science_en#open-science-practices (2024).

66. European Parliament and Council. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information. https://eur-lex.europa.eu/eli/dir/2019/1024/oj (2019). *Official Journal of the European Union*, L 172, 26 June 2019, pp. 56–83.

67. European Parliament and Council. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng (2019). *Official Journal of the European Union*, 17 April 2019.

68. Lowy Institute. Global Diplomacy Index. https://globaldiplomacyindex.lowyinstitute.org/ (2024).

69. Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. Bag of tricks for efficient text classification. *arXiv preprint*. doi:10.48550/arXiv.1607.01759 (2016).

70. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H. & Mikolov, T. FastText.zip: Compressing text classification models. *arXiv preprint*. doi:10.48550/arXiv.1612.03651 (2016).

71. Han, S. Googletrans 4.0.2. https://pypi.org/project/googletrans/ (2025).

72. Lu, Qiuhao, Li, Rui, Wen, Andrew, Wang, Jinlian, Wang, Liwei & Liu, Hongfang Large language models struggle in token-level clinical named entity recognition. *arXiv preprint*. doi:10.48550/arXiv.2407.00731 (2024).

73. Mikhaylov, S., Laver, M. & Benoit, K. R. Coder reliability and misclassification in the human coding of party manifestos. *Polit. Analysis* **20**(1), 78–91. doi:10.1093/pan/mpr047 (2012).

74. Deleger, L., Li, Q., Lingren, T., Kaiser, M., Molnar, K., Stoutenborough, L., Kouril, M., Marsolo, K. & Solti, I. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, vol. 2012, pp. 144–153 (2012).

75. Brandsen, A, Verberne, S, Wansleeben, M & Lambers, K. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4573–4577 (2020).

76. Li, D., Rosé, C., Yuan, A. & Zhou, C. Estimating agreement by chance for sequence annotation. *arXiv preprint*. doi:10.48550/arXiv.2407.11371 (2024).

77. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D. & *et al.* DeepSeek-V3 Technical Report. *arXiv preprint*. doi:10.48550/arXiv.2412.19437 (2024).

78. Zhan, Z., Zhou, S., Zhou, H., Deng, J., Hou, Y., Yeung, J. & Zhang, R. An evaluation of DeepSeek models in biomedical natural language processing. *arXiv preprint*. doi:10.48550/arXiv.2503.00624 (2025).

79. MediaWiki. API:Search. https://www.mediawiki.org/wiki/API:Search (2024).

80. Varghese, R & Sambath, M. YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 1–6 (IEEE, 2024).

81. Rizvandwiki. Gender-classification. https://huggingface.co/rizvandwiki/gender-classification/ (2023).

82. Muğurtay, N., Sirin, K.G., Najafabad, H.M, Kahya, A.T., Yılmaz, G., Zouzou, Y., Bahceci, B., Demir, A., Tosun, D., Bac, M.M., and Varol, O. GlobalDiplomacyNet Dataset. Harvard Dataverse. doi:10.7910/DVN/HYJDE0 (2025).

83. Panke, D. & Starkmann, A. Towards an increasing regionalization of international politics? An analysis of external competencies of regional international organizations. *Glob. Aff.* **7**, 43–65. doi:10.1080/23340460.2021.1913435 (2021).

DATA COLLECTION

Executive and Ministry of Foreign Affairs for **156** countries and **248** websites

Web Scraping

Selenium, Requests, VPN services, Internet Web Archive

1.16M+ webpages
1.18M+ images

ANALYSIS

Named-Entity Recognition

- Person
- Location
- Organization

Computational Image Analysis

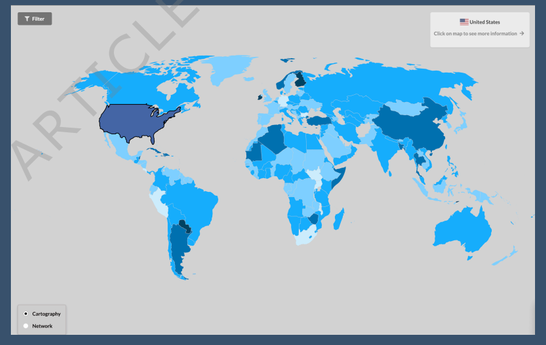
Scene details and estimates on people count, age and gender

Network & Content Analysis

- Country and people co-occurrence networks
- Topics and events
- Temporal changes

DATASET

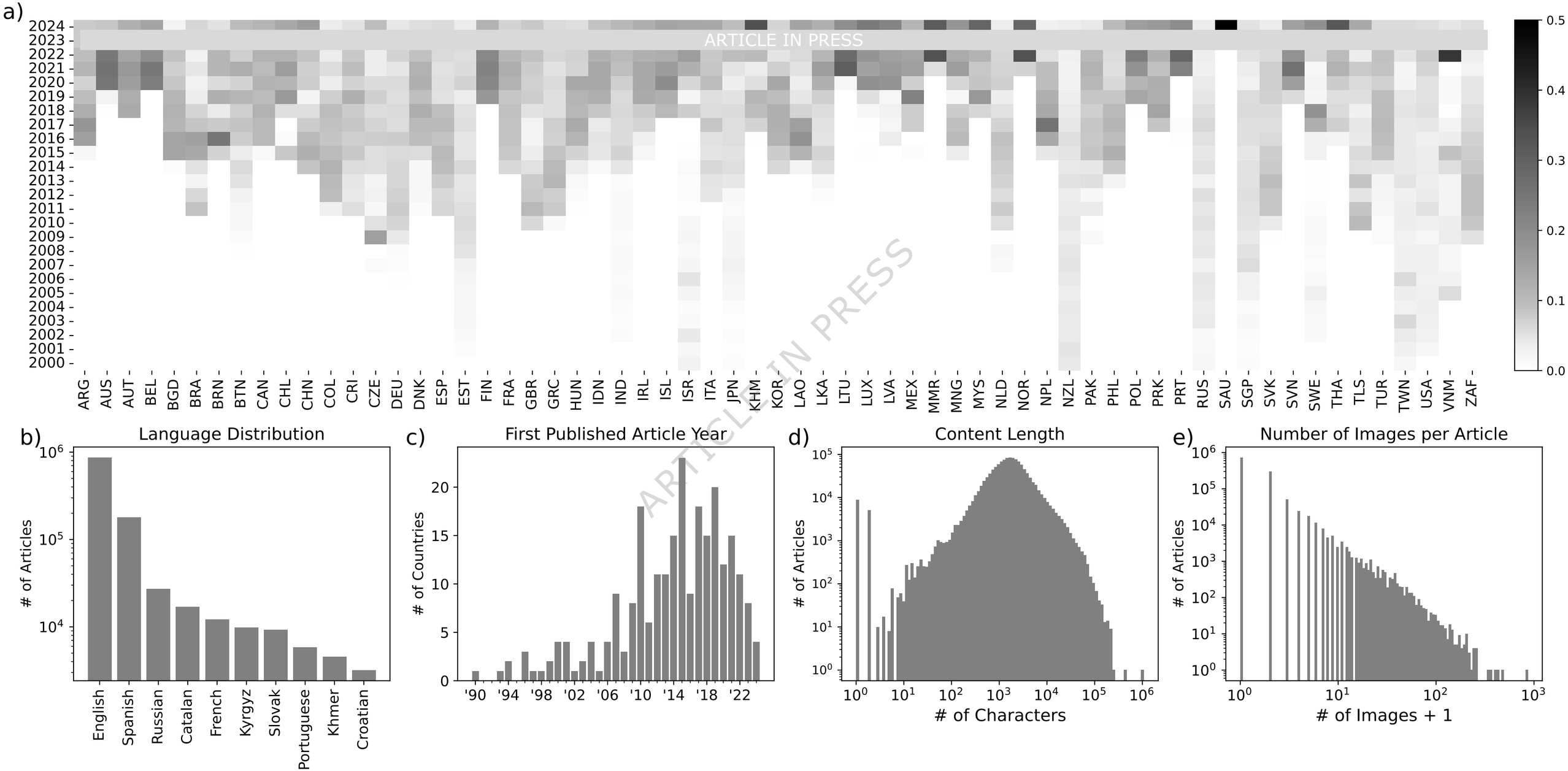
global diplomaticNet.

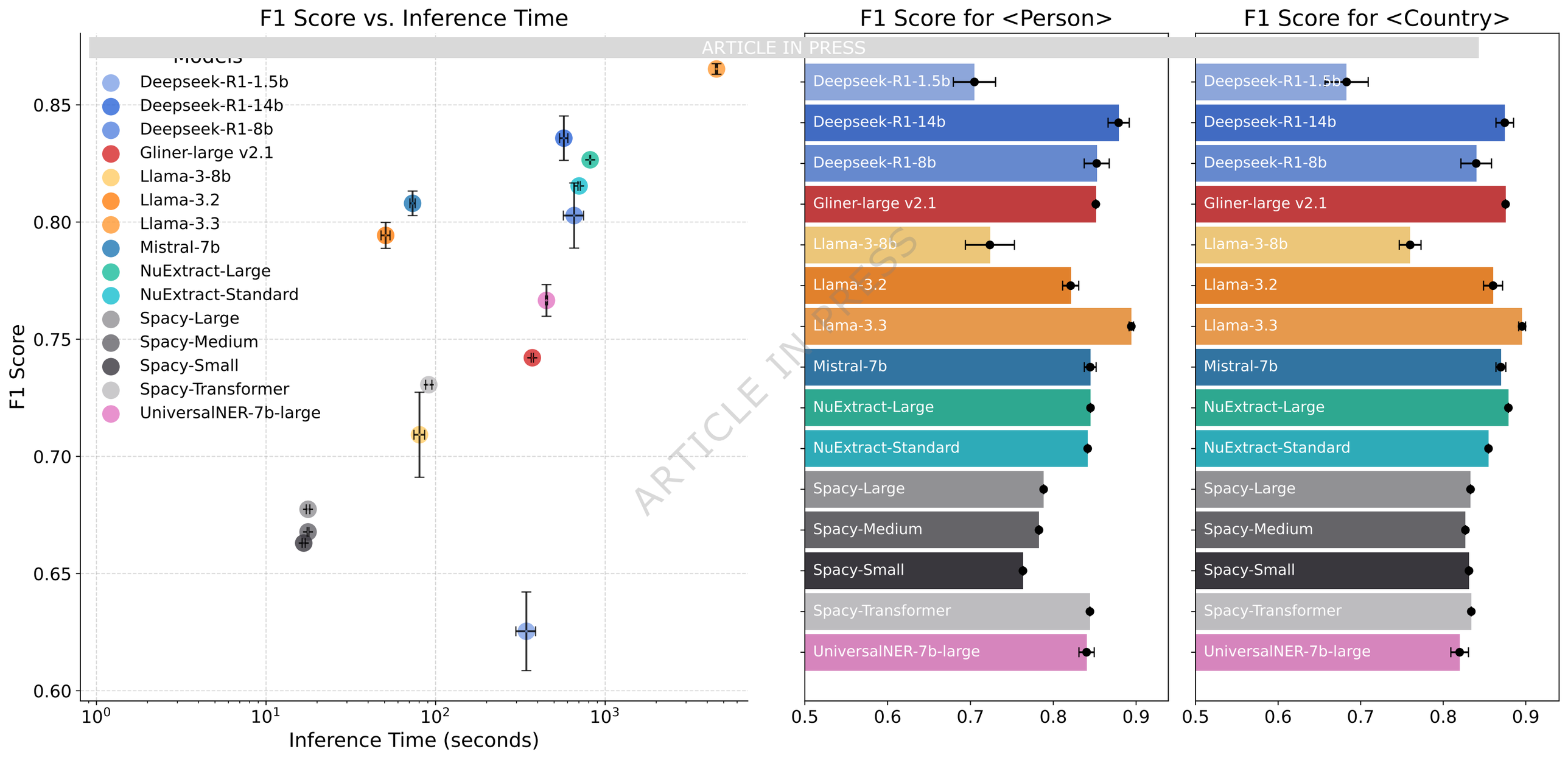


The Dataverse Project

Interactive Project Website → https://globaldiplomacy.net

Public Data Repository





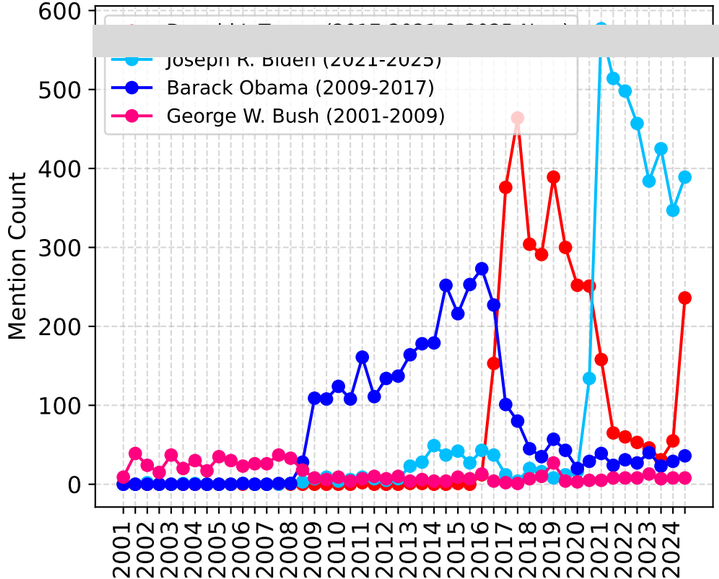
a) news.jsonl


```
{
  "id": "bacfe48df5be372e9ff880785e5e4bcf",
  "url": "https://www.tccb.gov.tr/en/news/542/91796/turkey-has-th...",
  "date": "2018-03-16",
  "title": "“Turkey has the power to take action for its own secu...",
  "content": " Speaking at the AK Party’s provincial congress in ...",
  "lang": "en",
  "title_original": "",
  "content_original": "",
  "entities": {
    "persons": ["Recep Tayyip Erdoğan","Mr. Obama","President Er..."],
    "countries": ["United States","U.S.","Iraq","Turkey","Syria"],
    "organizations": ["United Nations"]
  },
  "wikidata_qids": {
    "persons": ["Q39259","Q76","Q39259"],
    "countries": ["Q30","Q30","Q796","Q43","Q858"],
    "organizations": ["Q1065"]
  }
}
```

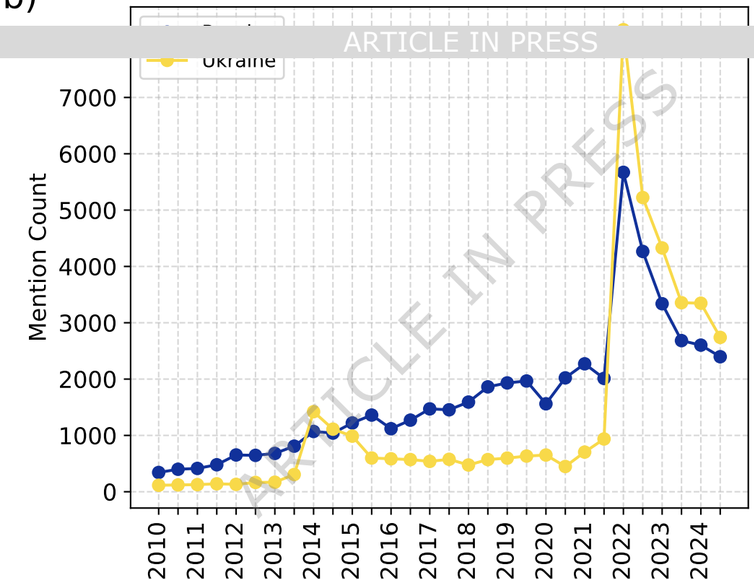

b) images.jsonl


```
{
  "id": "d963463741063cd5a4899172fc689fba",
  "news-id": "bacfe48df5be372e9ff880785e5e4bcf",
  "url": "https://www.tccb.gov.tr/ImageResizer/CropImage?w=-1&h=-...",
  "male-count": 6,
  "female-count": 2
}
```

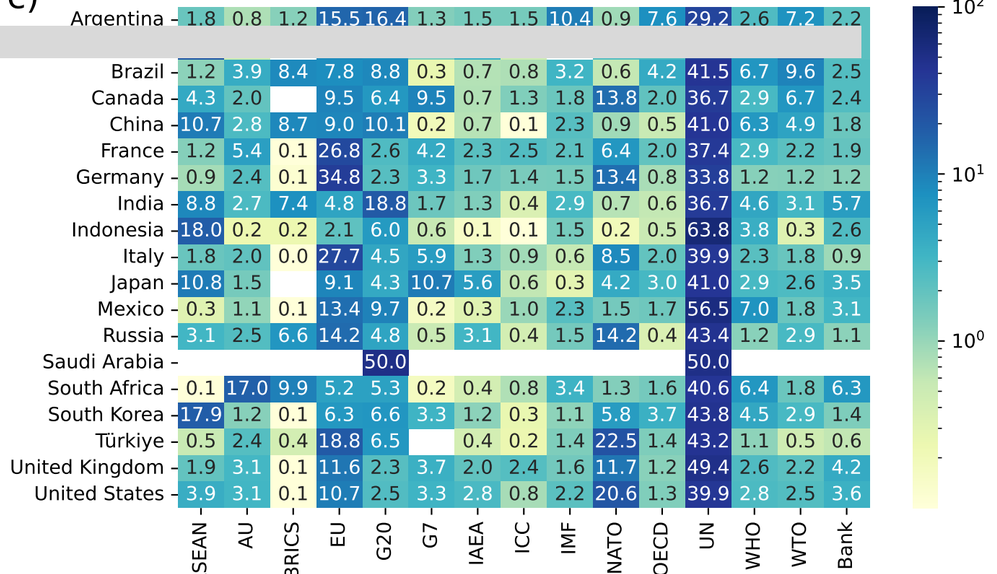
a)

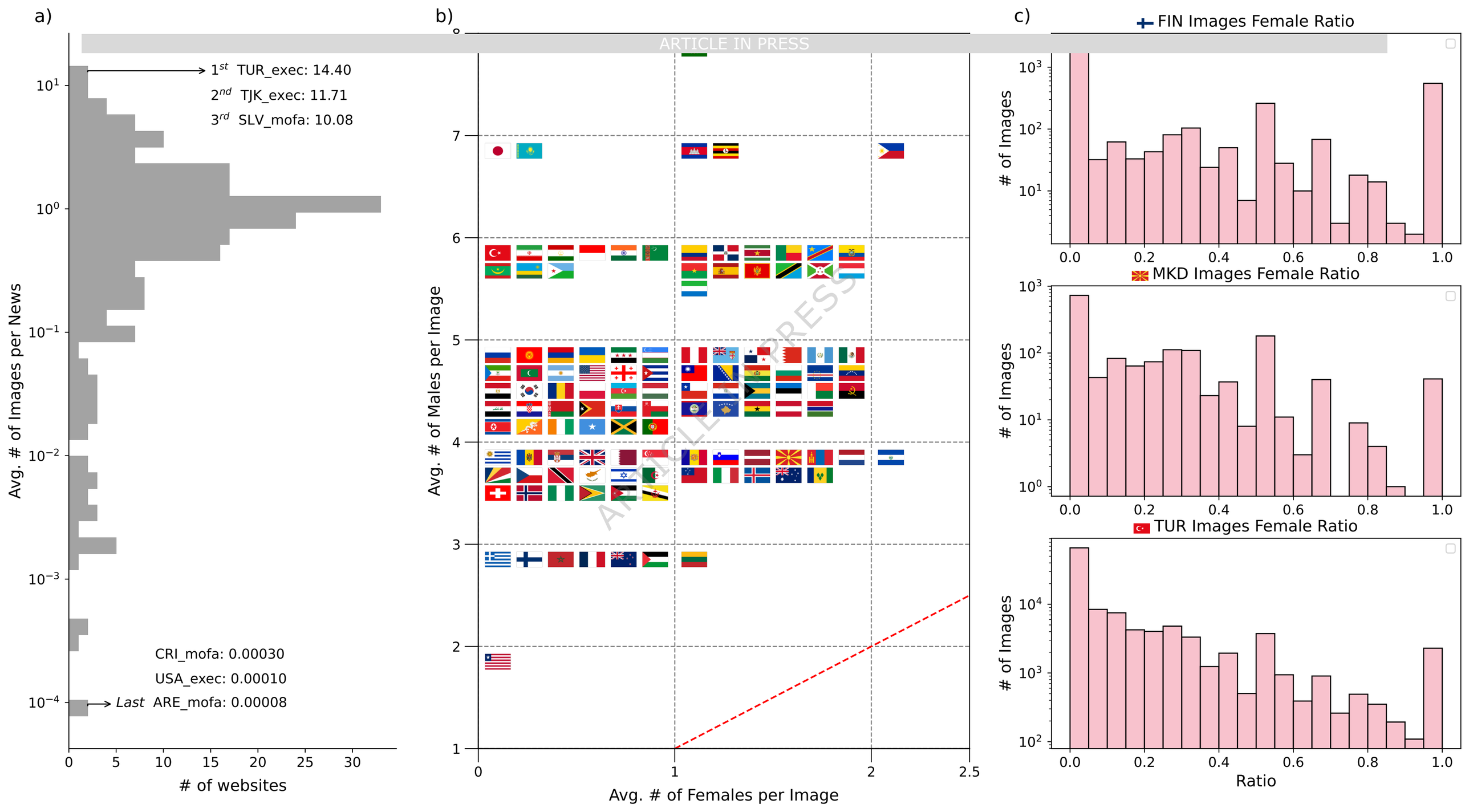


b)

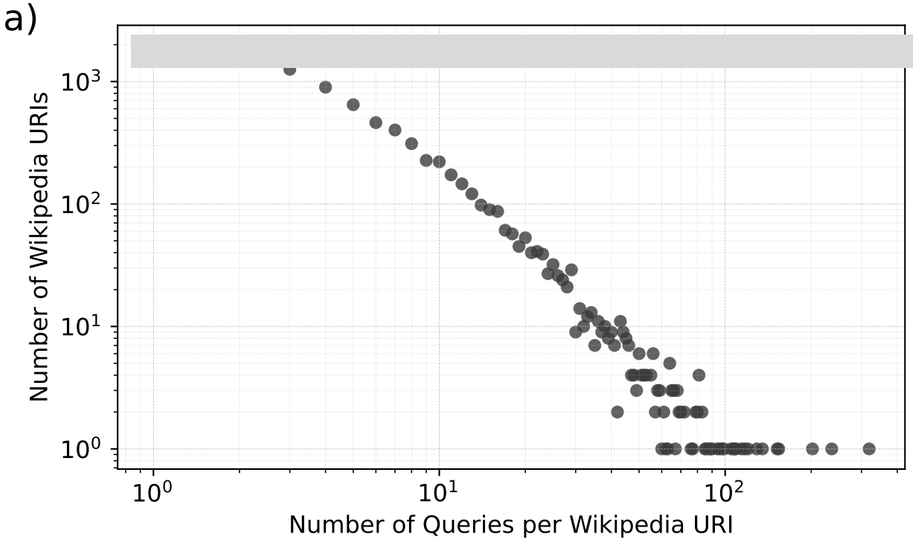


c)

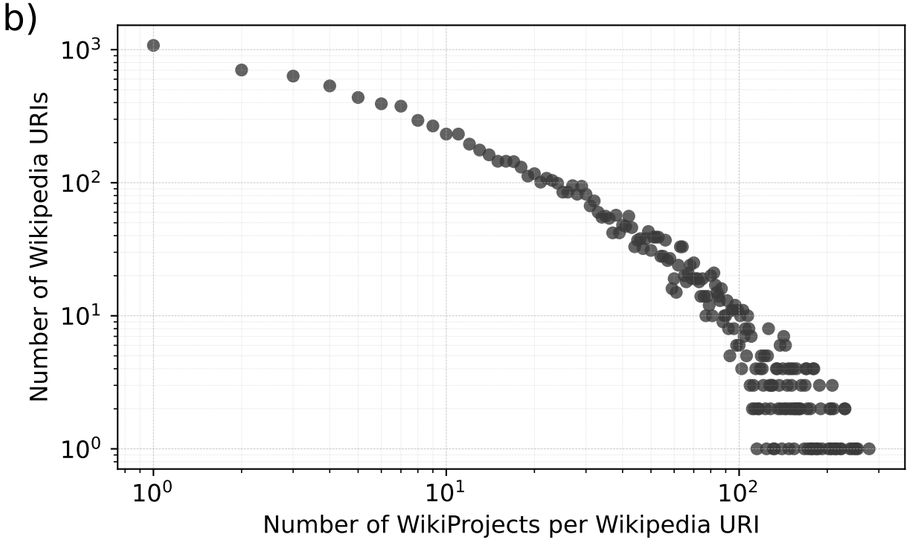




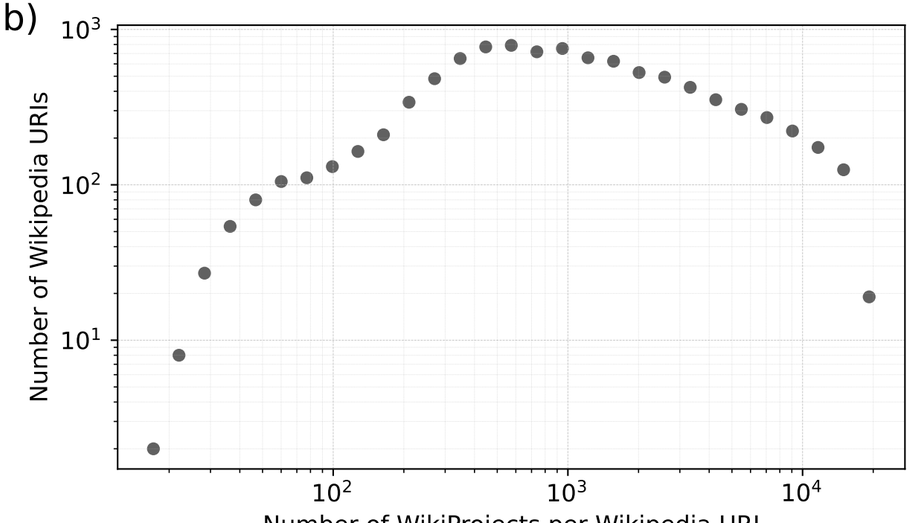
a)



b)



b)



d)