15

A CASE STUDY ON AI USAGE FOR COLLECTING PHILANTHROPY DATA IN THE WESTERN BALKANS

Nikola Milinković and Marko Galjak

1 Introduction

In an era where data reigns supreme, understanding the intricate web of philanthropy becomes both a challenge and a necessity. Philanthropic gestures, rooted deep within cultural, socio-economic, and political spheres, often echo the rich tapestry of diverse human motivations, aspirations, and needs. The Western Balkans, a region steeped in history and multifaceted identities, is a testament to the complexity of such philanthropic dynamics. Drawing connections and understanding giving patterns here is not merely an academic exercise but a pursuit that can inform, guide, and inspire impactful and sustainable philanthropic initiatives in a rapidly changing world.

The Giving Balkans philanthropy database emerges as a beacon in this quest, capturing the philanthropic heartbeat of seven unique Western Balkan countries: Serbia, Croatia, Bosnia and Herzegovina, Albania, Macedonia, Kosovo, and Montenegro. Documenting a staggering 90,313 philanthropic gestures involving 35,779 distinct entities and transcending half a billion euros between 2015 and 2022, the database serves as a record and a mirror reflecting the nuanced dance of giving in the region. As we delve deeper into this chapter, we aim to unravel the digital transformation journey of this database, accentuated by the adoption of artificial intelligence. Through a case study lens, we will shed light on the successes, challenges, and the forward trajectory of integrating AI for better, more efficient, and more insightful data collection and analysis.

This chapter is written in the middle of the transition to AI. It offers a snapshot of the current state of using AI for philanthropy data gathering and processing in the Western Balkans and is in no way the final form. With the field of AI evolving quickly, this is likely the case in two years. Our process will evolve following the rapid advancement of AI methods and tools.

1.1 The Western Balkans' rugged philanthropic landscape—a contextual glimpse

The philanthropic landscape of the Western Balkans is complex. The first step in grasping the complexity is understanding the context in which Giving Balkans philanthropy data is collected. The Western Balkans is an intriguing ensemble of seven nations presenting a rich mosaic of shared histories and unique trajectories. While the majority of the countries—Croatia, Bosnia and

DOI: 10.4324/9781003468615-18

Herzegovina, Montenegro, and Serbia—were once tethered under the banner of Yugoslavia and share linguistic ties, Albania is different, marked by isolation during the Hoxha regime (Glenny, 2001). Although retaining its linguistic link with Albania, Kosovo traversed its journey alongside the former Yugoslav states. The linguistic diversities, with Macedonia's South Slavic language and the Albanian threads of Kosovo and Albania, add another layer to this complexity (Duncan, 2016). These linguistic differences pose a significant challenge in collecting philanthropy data consistently across the region. The challenge is primarily in terms of human resources. Overcoming the linguistic barriers requires a skilled workforce fluent in the region's diverse languages.

Moreover, an in-depth understanding of the local civic and philanthropic landscapes is important, often necessitating the recruitment of knowledgeable personnel from within each country. This ensures that data collection is linguistically accurate and culturally and contextually informed, an important aspect for reliable and comprehensive philanthropic data gathering. Another consideration is the verification process, where Philanthropy Data Analysts check with either donor or beneficiary about the donation instance that has appeared in the media. Having a local inquiring about veracity and additional information is much more likely to solicit a response than reaching out from a different country. This can be an important consideration in light of the region's historically sensitive context.

There are also considerable differences in the economies of the Western Balkans, as the poorest country (Kosovo) has a gross domestic product per capita (GDPPC) of $5,531.5. In contrast, the richest, Croatia, has a GDPPC of $18,413 (World Bank, 2023). This more than threefold difference in countries' economic output translates to different philanthropic contexts (i.e., how philanthropy is practiced and therefore tracked can also be different). This layer adds additional methodological difficulties.

The post-socialist tapestry these nations wear resonates with overtones of state dependency rooted in the shared communist history. Citizens, influenced by this past, often look toward the government as the primary steward of societal welfare (Grødeland, 2006). This outlook on governance, combined with the region's sporadic political instabilities (EWB, 2023), corruption challenges (Transparency International, 2020), and evolving democracies (Freedom House, 2016), intricately shapes the philanthropic motivations and actions here. In this region, the concept of philanthropy often diverges from traditional forms like fundraising for community development or crowdfunding for community projects; such practices are not deeply ingrained. Instead, a significant amount of philanthropy is informal and immeasurable, characterized by people spontaneously helping each other in times of need. Civil Society Organizations (CSOs) have become heavily reliant on an influx of foreign funding, a trend that began in the 1990s. Recently, a notable shift in philanthropic focus has emerged, with crowdfunding campaigns for healthcare, especially for sick children, gaining prominence and attracting considerable support over the last decade. The ingrained perception of the state as the primary provider of all societal needs, juxtaposed with the emergent necessity to crowdfund for the immediate healthcare needs of sick children, underscores a significant shift in the region's philanthropic landscape. This juxtaposition highlights a changing dynamic where the public increasingly recognizes the limitations of state support and turns toward philanthropy for urgent and critical needs.

The legal environment is a pivotal factor in the philanthropic scenario of the Western Balkans. While offering growth avenues for non-profits, the legislative matrix also interposes specific barriers that might constrict their impact and reach (USAID, 2023) (e.g., difficulties around regulations of volunteer work and lack of legal frameworks that support volunteerism, no tax incentives for philanthropy, collecting VAT on food donations, etc.). Additionally, the region's socio-economic fabric, a product of its transition from socialism to a more economically and politically liberal

regime, plays a significant role in how philanthropy evolves and operates. Economic transition brought with it new wealth and economic disparities. With no legacy philanthropic organizations, a nascent CSO sector, the strong influence of international donors, and clumsy corporate responsibility adopted from multinational corporations' headquarters, philanthropy in the Western Balkans evolved uniquely in the past three decades.

The difficult economic situation and wars boosted traditional emigration from the region to the more prosperous countries. While diaspora communities might be deeply integrated into their host countries, spanning generations, they often maintain a pronounced inclination toward philanthropic efforts directed at their countries of origin (Brinkerhoff, 2014). This highlights the potential for viewing the diaspora as a local resource for philanthropic endeavors. This is especially true in countries like Bosnia and Herzegovina, Albania, and Kosovo, where remittances form a significant economic component (Bajra, 2021), and the diaspora's role becomes central. As the Western Balkans grapple with the diminishing influence of foreign donors, cultivating a robust, locally sourced philanthropy ecosystem emerges as a crucial fulcrum for both the sustenance of non-profits and the democratic fabric of the region.

The Western Balkans' philanthropic landscape is shaped by several key factors: diverse linguistic backgrounds complicating data collection, significant economic disparities impacting philanthropic practices, a post-socialist legacy influencing public reliance on government for welfare, and legal challenges that restrict non-profit activities. Additionally, the socio-economic shift from socialism and the unique role of the diaspora, especially in countries heavily reliant on remittances, significantly shape the region's approach to philanthropy. These elements collectively contribute to the complexity of philanthropy in the Western Balkans.

1.2 Catalyst Balkans—the organization behind the Giving Balkans

Catalyst is a Serbian-registered foundation launched in early 2013 to promote the growth and improved transparency of individual and corporate philanthropic culture in the Western Balkans and to further the digital transformation of the non-profit sector. Based in Belgrade, Serbia, Catalyst covers seven countries.[1] Catalyst Balkans has built its reputation as a go-to partner for information, support, or advice on domestic giving by taking a systems-based approach to broaden and deepen the Western Balkans' philanthropy ecosystem. It is an example of locally led development; Catalyst was founded specifically to address gaps in the ecosystem and work in partnership with established and emerging stakeholders. Catalyst's active participation in ecosystem entities includes the Serbian Philanthropy Forum, Philanthropy Forum of Bosnia and Herzegovina, Kosovo Philanthropy Forum, Southeastern Europe Indigenous Grantmakers Network (SIGN) Network, and the European Research Network on Philanthropy (ERNOP).

Catalyst Balkans provides tech and services to the philanthropy and non-profit ecosystems of the Western Balkans. Through Donacije.rs,[2] 198 non-profits have raised $1.18 million. Using CiviCatalyst.org, a service based on open-source CRM (Constituent Relationship Management) software for non-profits that Catalyst provides hosting and customization services, 120 non-profits manage their data more securely. Three hundred and fifty Serbian non-profits have claimed a transparency badge on Neprofitne.rs. The unique Giving Balkans methodology for collecting transaction-level micro data on domestic philanthropic flows in seven Western Balkans countries relies on a combination of press clipping and direct verification of gathered data with recipients and donors. Using this data, Catalyst identifies several key trends in philanthropy. Firstly, there is an increasing tendency toward mass individual giving, indicating a shift in how individuals

contribute to charitable causes. Secondly, there is a notable change in how corporate donors operate. Corporations are becoming more sophisticated in their giving strategies, often channeling their donations through non-profits, even when public institutions are the intended final beneficiaries. This approach signifies a strategic move toward leveraging the expertise and networks of non-profits for more impactful giving. Lastly, there is a rise in community-based philanthropy, reflecting a growing emphasis on localized, grassroots efforts to address societal needs.

1.3 Giving Balkans database

Giving Balkans gathers data on charitable giving in the Western Balkans region using alternative methods, primarily sourcing information from media reports and other readily available resources. Official data on philanthropy from key institutions such as the Ministries of Finance and the Tax Administration is absent in the region. To bridge this gap, Giving Balkans continuously monitors printed, electronic, and online media on local, regional, and national levels within the Western Balkans for any instances of giving. This data can then be easily accessed by non-profit organizations, corporations, and individuals, providing a better macro understanding of the entire ecosystem and concrete philanthropy intelligence.

As of late 2023, the database contains more than 87,000 instances of giving by more than 13,000 distinct donors to more than 24,000 distinct beneficiaries, amounting to more than 644 million euros (Catalyst Balkans, 2023).

Giving Balkans app is an interactive web application built using the R programming language (R Core Team, 2023) and Shiny (Chang et al., 2023) R package, both of which are open-source (see Figure 15.1). The user-friendly app facilitates data exploration through intuitive filtering across all dimensions by clicking on the visualizations. This ease of use empowers those without a technical background to delve into what is often called "philanthropy intelligence." Such insights can assist non-profit organizations in strategizing their fundraising initiatives and, equally, guide donors in identifying the non-profits they wish to collaborate with.

The data from Giving Balkans is rich and relationally structured, paving the way for graph creation. A unique feature incorporated into the Giving Balkans app is CiviGraph. This tool empowers users to navigate the intricate philanthropy networks built around specific entities in the database (Galjak, 2020). For instance, users can investigate which donors contributed to particular organizations while simultaneously viewing the donors' immediate philanthropic neighborhood. This involves understanding the other organizations a donor has contributed to and identifying other donors for these organizations. The ability to access and analyze this data offers significant strategic value. Organizations can leverage these insights to identify potential partnerships, optimize fundraising strategies, and better understand the dynamics of the philanthropic landscape. Beyond its strategic value for organizations seeking partnerships and optimizing fundraising strategies, CiviGraph is also a potent tool for investigative journalism tracking donations from politically significant entities, like companies with local or foreign government stakes such as Russian or Chinese, and mapping their philanthropic impact in the region (see Figure 15.2).

1.4 Why was Giving Balkans created?

The primary motivation behind Giving Balkans was to illuminate the landscape of philanthropy, given the absence of other comprehensive data sources. In the Western Balkans, the regulatory and tax frameworks do not capture any significant data about charitable giving. The only other

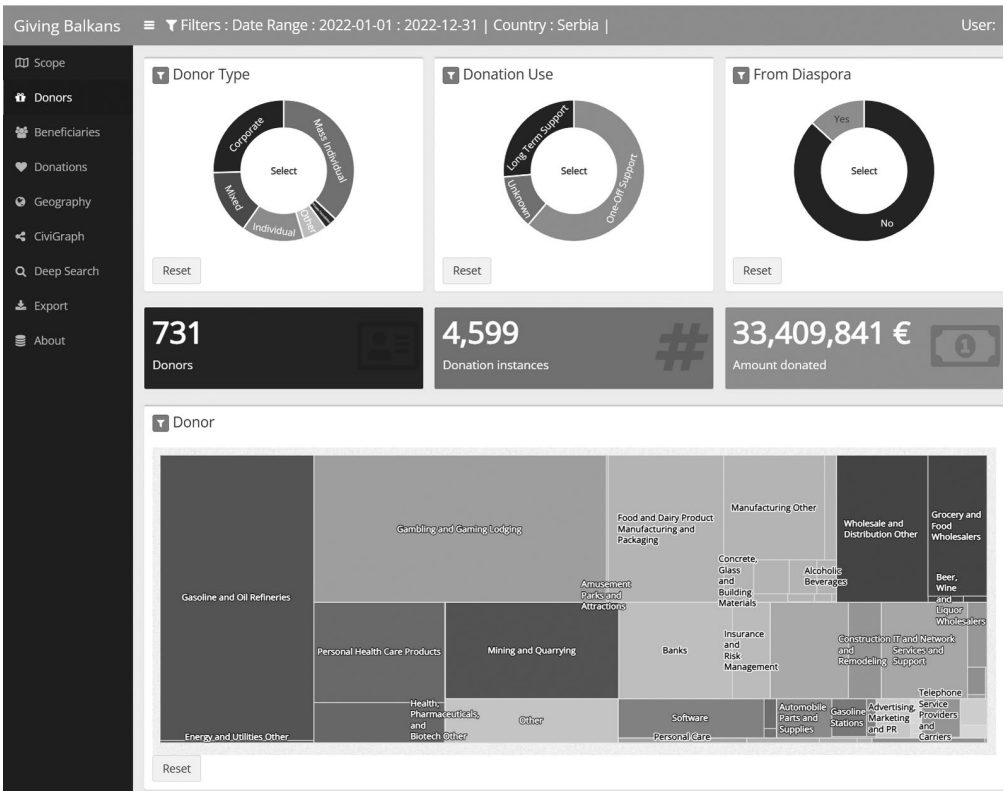

Figure 15.1 Giving Balkans interactive data visualization and analysis web application.

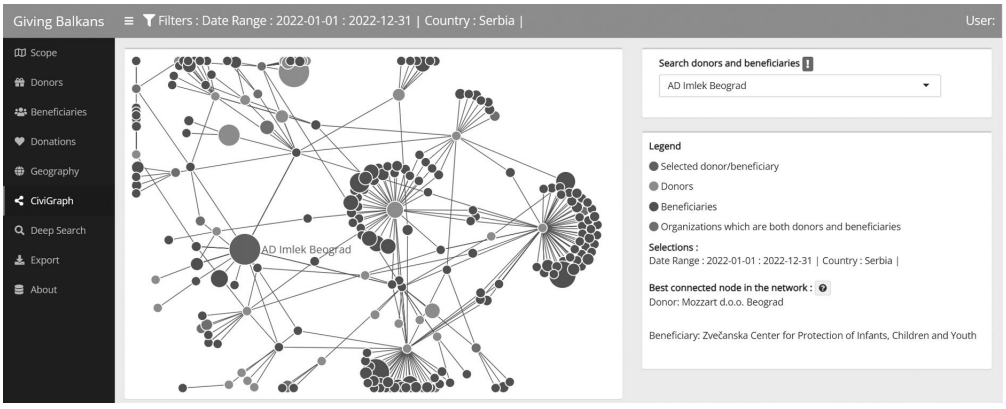

Figure 15.2 CiviGraph—a social network analysis tool built into the Giving Balkans app leveraging the relational data.

glimpses into the region's philanthropic activities came from sporadic ad hoc surveys. There was a prevailing sentiment in the region that philanthropy was either minimal or non-existent and could not serve as a funding source for CSOs. These CSOs have depended on foreign funds to sustain their operations for years. Thus, Giving Balkans primarily aimed to debunk this notion by demonstrating the existence of substantial local resources that CSOs, among other entities, can harness. Beyond presenting a macro perspective of the philanthropic ecosystem, Giving Balkans is important in providing a micro view, offering detailed data points that can be instrumental in optimizing philanthropic giving.

The significance of readily available data became especially evident during the COVID-19 pandemic. The pandemic notably impacted the Western Balkans region (Marinković & Galjak, 2021). Civil society organizations in the region have faced sustainability challenges due to the economic and social disruptions caused by the pandemic (Drobarov et al., 2021). Catalyst Balkans (2020) sprung into action in response to these challenges, assisting 26 different non-profits in launching fundraising campaigns. Leveraging the Giving Balkans database, they helped these non-profits in crowdfunding efforts and locating corporate donors, ultimately raising over 200,000 euros. Throughout the COVID-19 crisis, Catalyst Balkans diligently monitored philanthropic activities in the Western Balkans. Their observations underscored significant contributions, predominantly directed toward essential supplies, and also captured the diverse donor dynamics across various countries (Catalyst Balkans, 2021). Such invaluable data can be harnessed in future crises to identify responsive donors who have previously demonstrated a readiness to contribute promptly.

1.5 Original process of data collection and methodology

Since its inception in 2013, Giving Balkans has primarily sourced its data from various media outlets, including newspapers, internet portals, television, and radio. Fortuitously, a company specializing in keyword press clipping services was available to cater to all seven countries covered by Giving Balkans. As illustrated in Figure 15.3, the initial methodology required human intervention. Each country's designated Philanthropy Data Analyst would manually process the press clipping data. This press clipping service would consistently forward media records containing predefined keywords (or combinations thereof) set by Catalyst Balkans for each language. These records would then undergo thorough processing by the Philanthropy Data Analysts (Galjak, 2020).

1.5.1 The problem of actual relevance for Giving Balkans database

The reliance on a keyword-based approach inevitably led to the inclusion of numerous false positives, which were not pertinent to the Giving Balkans database. For instance, news of a US-based celebrity donating to a charity in an African nation might dominate media outlets, ticking all the keyword boxes. However, such a story is not relevant to the Giving Balkans records. A more specific example of this challenge is the 2022 Russian invasion of Ukraine. The coverage around military donations made to Ukraine by various countries triggered the designated keywords, but these stories held no relevance to the Giving Balkans database.

When evaluating a news item, the Philanthropy Data Analyst must discern whether the article genuinely pertains to the Giving Balkans database. This judgment hinges on the concept of "local philanthropy." The origin of the donor characterizes local philanthropy. Suppose the donor hails from a country within the Western Balkans or belongs to the diaspora of one of the Western Balkans countries, and the donation is intended for a beneficiary in their country of origin. In that case, it is classified as local philanthropy.

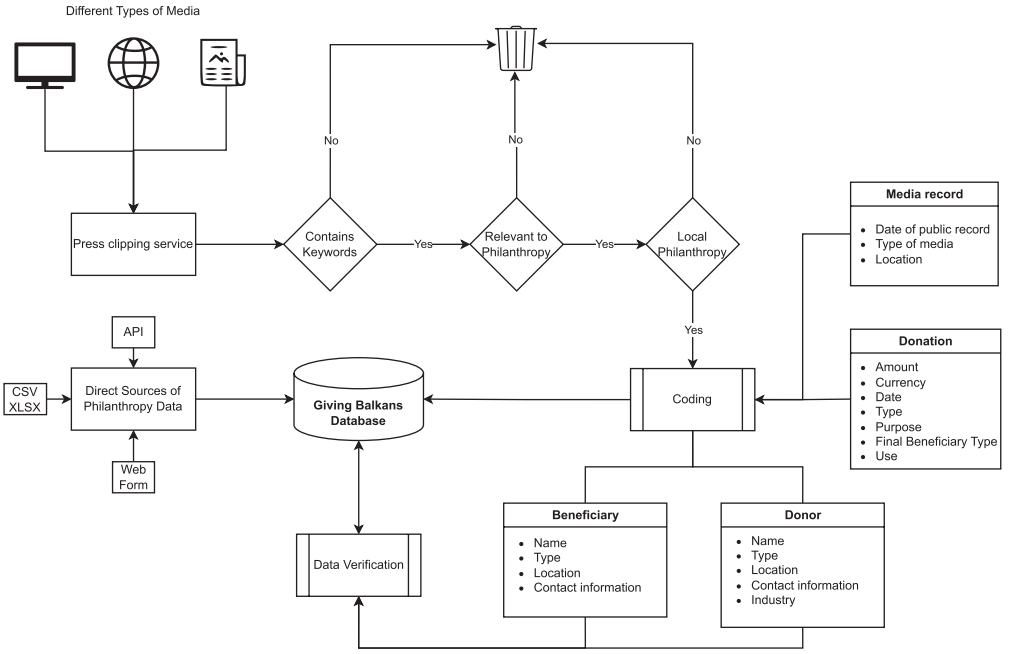

Figure 15.3 The process of collecting data for Giving Balkans database.

1.5.2 The problem of languages

As delineated in the introduction, our staff processes media articles in several distinct languages: Serbian, Croatian, Bosnian, and Montenegrin (which are essentially variations of the same language), Albanian, and Macedonian. Consequently, the staff responsible for each country must be proficient in one or more of these languages. This is particularly crucial for Macedonia, which has a significant Albanian minority. Such linguistic requirements make the recruitment of Philanthropy Data Analysts challenging. The work structure at Catalyst Balkans has consistently been organized by language. For instance, an Albanian-speaking staff member might be responsible for covering Albania, Macedonia, or Kosovo.

1.5.3 The problem of categories

When media-sourced information is deemed pertinent, it is coded into the database by the Philanthropy Data Analysts. This encoding procedure encompasses several stages. Paramount among these is the categorization of the involved entities (both donor and beneficiary) and the specific instance of the donation. The details about these entities often necessitate additional research, as media articles might not furnish comprehensive data. For instance, an article may state that a local company donated to a neighborhood charity. In such cases, the analyst must ascertain the donor's type (from 11 possible options), identify their industry (from a list of 157 possibilities), and gather relevant contact details, which include address, phone number, email, website, and social media accounts. Similarly, for the beneficiary, the analyst is tasked with pinpointing the type of beneficiary (choosing from 13 options) and documenting analogous contact particulars

as those noted for donors. The donation itself can be categorized in multiple ways, including the category of the donation instance (with ten choices), the type of donation (six options), the purpose of the donation (spanning 26 options), and the category of the end beneficiaries (from 37 available choices). Excluding geographical considerations, like selecting the municipality of the donor and the beneficiary, results in over a billion potential combinations for each donation record. To ensure accuracy and integrity, the Data Quality Manager supervises the whole process, and verifies that all information is correctly coded.

1.5.4 The problem of duplicates

True originality in news is uncommon, especially given that when a story breaks in one media outlet, it is frequently replicated across many others—a trend particularly evident with online media. This replication introduces challenges in volume; articles, though essentially echoing the same information, often bear distinct, clickbait-inspired titles. This means that those responsible for processing the data frequently find themselves navigating through numerous articles that, content-wise, are virtually identical.

1.5.5 The problem of truth

Not everything reported in the media is accurate. Hence, verification becomes a pivotal task for Philanthropy Data Analysts. Every recorded donation is cross-checked with at least one of the involved parties, be it the donor or the beneficiary. This ensures that the information we have sourced from media outlets aligns with the facts. At times, this verification leads to updates in our records, be it regarding the donation amount or other details that weren't initially covered in the media reports. Approximately two-thirds of the donations are validated through this verification process, leaving a third unverified.

1.5.6 Data harvesting

Besides the routine press clipping for data collection, Catalyst Balkans also directly harvests data from available sources. This includes data from Donacije.rs, which, while managed by Catalyst Balkans, only represents a minor portion of the overall charitable giving in Serbia. On the other end of the spectrum, we have direct API access to data from Budi Human (Serbian for *be humane*), an organization dedicated to fundraising for individuals with health challenges, accounting for a significant portion of total donations in Serbia. Additionally, certain companies opt to provide their donation data directly to Catalyst Balkans via email or web forms integrated into the Giving Balkans website.

1.6 Limitations of the methodology

Given the lack of consistent sources for assessing charitable donations in the Western Balkans, Catalyst Balkans has adopted innovative data collection methods. These are primarily based on print, online, and electronic media and are supplemented by other available data platforms. However, this approach has limitations: not all philanthropic actions are highlighted in the media, and the published reports often lack the necessary details to understand philanthropy trends fully. Beyond these media-centric strategies, the Giving Balkans database uses direct data channels. Some organizations, for instance, provide firsthand access to their donation data via Application

Programming Interfaces (APIs) or by regularly sharing spreadsheet files. While this direct method simplifies data integration and bolsters accuracy, it comes with challenges, like reliance on third parties and potential inconsistencies in data. While our data might not capture the entire landscape, it does establish baseline figures, indicating the minimum number of events, financial contributions, and participants recorded annually. These figures offer a foundational perspective on the basic level of philanthropic activities in a country. One of the main challenges is tracking the growth of philanthropy in an environment with sporadic data collection and inconsistent examination. To tackle this, Catalyst Balkans has developed a set of preliminary criteria to shed light on the various aspects of charitable donations. These cover charitable events or drives, financial collection methods, guiding donation principles, recipients and beneficiaries of donations, the donors, stakeholders, and the extent of media coverage. Currently, quantitative and qualitative metrics linked to each criterion offer a solid framework for assessing the philanthropic terrain of a nation over several years. Regarding data reliability, the Giving Balkans database is updated daily, reflecting a consistent commitment to accuracy and timeliness.

2 AI-assisted process of data collection

Ever since 2017, Catalyst Balkans has looked for ways to automate some of its processes around collecting and processing philanthropy data. Replacing Philanthropy Data Analysts seemed like an impossible task. Increasing analysis processing costs and the volume of press clipping data finally pushed us to develop the AI-assisted data collection and analysis process. The idea is to maximize Philanthropy Data Analysts' productivity instead of creating fully autonomous agents that would replace them. The question was how to achieve this: by addressing the two major problems—false positives and duplicate articles (see details in the Resulting Solution section).

To this end, we have created a system for preprocessing articles that solves these problems and significantly increases Philanthropy Data Analysts' productivity (Figure 15.4). Our system utilizes custom, language-specific models instead of relying on pre-trained multilingual models (such as RoBERTa). We initially opted for xlm-roberta-base, a general model that needed to be fine-tuned for the downstream task of text classification. It seemed perfect as it covered all the languages of the Western Balkans. However, we have failed to make it work for this task using our train and test datasets in Serbian language only. An additional obstacle was that we needed to handle this fine-tuning on the Azure cloud platform, which presented additional costs. We opted for custom-trained language-specific models after initial testing showed very promising results. However, we acknowledge the potential long-term benefits of fine-tuning multilingual pre-trained models. Investing more resources into this approach in the future could yield a more robust and scalable solution suitable for a broader range of languages and tasks. Therefore, revisiting and refining the strategy of utilizing multilingual pre-trained models remains a consideration for our future work.

2.1 Text processing

In our data collection and analysis process, we initially extracted data in batches of a thousand articles. For precision, every article undergoes an exhaustive, fully autonomous cleaning procedure consisting of two steps.

In the first step, poorly OCR-ed (optical character recognition) articles from print media are detected and flagged as such using PCA-based outlier detection on character-level n-gram (n = 1,2,3) based on Term Frequency–Inverse Document Frequency algorithm. This was necessary as some articles, particularly those digitized from print sources, might contain errors from imperfect

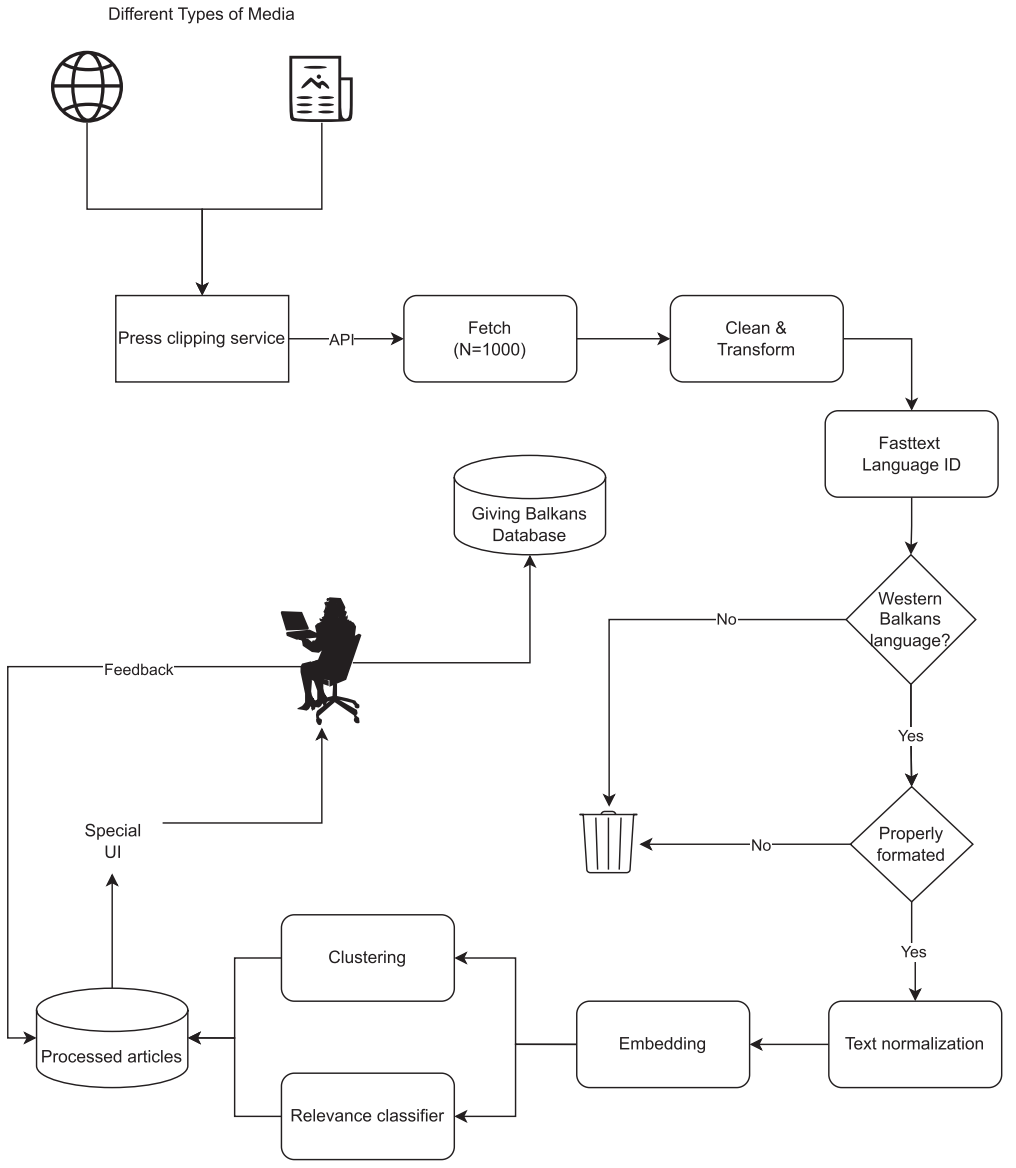

Figure 15.4 The AI-assisted process of collecting data for Giving Balkans database.

scanning techniques. Identifying and flagging such articles was required to ensure their errors did not skew our analysis.

In the second step, the text is normalized. First by Unicode Normalization Form: Compatibility (K) Composition, then articles in Serbian Cyrillic are transliterated into Latin, repeated horizontal whitespaces are eliminated, and line endings are normalized (“\r\n,” “\n\r,” “\r” to “\n,” etc.).

Once these adjustments are made, we structure each article into distinct paragraphs and sentences, facilitating smoother processing and analysis. We discovered that depending solely on the language

information from the source could be misleading. Hence, we employ a specialized model known as fastText (Joulin, Grave, Bojanowski, Douze, et al., 2016; Joulin, Grave, Bojanowski, & Mikolov, 2016) to accurately determine the article's language. Subsequently, we standardize each article's text.

2.2 Embeddings

In the expansive realm of AI-powered textual analysis, embeddings stand out as a refined tool for distilling the core meaning of documents and words by converting them into numerical vectors. This mathematical portrayal enables computers to identify patterns, themes, and resemblances across extensive datasets.

2.2.1 Adopting Doc2vec and Word2vec

To process the vast array of articles from the Western Balkans, we utilized the Doc2vec (Le & Mikolov, 2014) and Word2vec (Mikolov et al., 2013) models from the Gensim library (Rehurek & Sojka, 2011). These models are acclaimed for their effectiveness, even with limited hardware resources. Our goal was to represent each news article as a vector—a feature of Doc2vec. However, we needed to train Word2vec models for each respective language to establish the multidimensional vector space for these representations. We began by training our Word2vec models on comprehensive Wikipedia corpora for Serbian/Bosnian/Croatian/Montenegrin, Macedonian, and Albanian languages. With these broad-scope embeddings in hand, we then tailored our system using our specific collection of news articles from 2015 to 2021. This allowed us to train our Doc2vec models, making them particularly attuned to the context of philanthropy, as highlighted by Lau and Baldwin (2016).

2.2.2 Article processing and model parameters

Before training, articles were processed and standardized: they were separated by language, deduplicated, broken down into paragraphs, and cleansed of punctuation and numerals. Notably, stemming and lemmatization were not employed due to their negligible impact on our pilot evaluations. For the technically inclined, our Doc2vec training used the PV-DBOW model variant alongside simultaneous skip-gram Word2vec training, with a 15-word context window and a 300-dimension vector embedding. We cycled through this training for a comprehensive 100 epochs.

2.2.3 Topic-based clustering and modeling

Beyond basic embeddings, we integrated Top2vec (Angelov, 2020), a cutting-edge tool for topic modeling. By default, Top2vec offers robust results through an intuitive API, autonomously handling hyperparameters and determining distinct topic counts. With just a few lines of code, Top2vec can train document and word embeddings, reduce dimensionality using uniform manifold approximation and projection—UMAP (McInnes et al., 2018), cluster these reduced vectors, and determine topic vectors. This was invaluable for static collections. However, our continuous inflow of articles presented challenges. Our solution was to bypass Top2vec's default training and instead employ our pre-trained Doc2vec embeddings. This adaptation permitted us to cluster articles in digestible batches daily, maintaining a consistent semantic vector space. As we forge ahead, we aim to develop a mechanism that seamlessly interlinks these daily clusters, tracing the narrative arc of news topics over time.

2.2.4 Relevance classifier

To effectively process the vast influx of news articles, we needed an automated system to quickly identify which articles were related to philanthropy and which were not. This was paramount, as an overwhelming majority—almost two-thirds—of the articles we received had no direct connection to our area of interest.

To achieve this, we embarked on a meticulous two-month project. We collected a large number of articles in the Serbian/Bosnian/Croatian and Albanian languages. Each article was carefully labeled as either relevant to philanthropy or irrelevant. For instance, in the Serbian/Bosnian/Croatian dataset, out of 26,045 articles, 13,425 were deemed relevant, while 12,620 were deemed unrelated to our focus.

While the irrelevant articles were primarily identified and labeled manually, the relevant articles were more straightforward to gather. We did this by cross-referencing with our pre-existing database of donations.

To translate this labeling effort into an actionable system, we utilized specialized mathematical models—vector embeddings—from our Doc2vec models. Combined with our labeled articles, these embeddings allowed us to train and evaluate several methods to automatically classify incoming articles from the scikit-learn library (Pedregosa et al., 2011). After testing various methods, we found the most success with an algorithm known as the support vector machine classifier (Chang & Lin, 2011; Platt, 1999) with a polynomial kernel of degree 3. In evaluating the algorithm's performance in classifying relevant articles, it showed remarkable efficiency in various languages. For the Serbian/Bosnian/Croatian language, the classifier exhibited a precision of 96.0%, a recall of 95.4%, and an overall accuracy rate of 95.6%. Similarly, when applied to Albanian articles, the classifier achieved a precision of 99.0%, a recall rate of 91.4%, and an accuracy of 95.63%.

In simpler terms, our system became exceptionally adept at sorting through heaps of news articles and pinpointing those relevant to philanthropy, all while requiring minimal human intervention.

2.3 Resulting solution

The resultant solution was coupled with a tailored user interface designed specifically for these new AI-assisted functionalities. Collectively, these modifications spurred significant productivity enhancements. The issue of false positives, where Philanthropy Data Analysts were inundated with media articles unrelated to philanthropy, can now be promptly identified and labeled as irrelevant. This is especially beneficial since this data informs subsequent training phases. The essence of semantic clustering allows for bulk categorization of articles about the same topic, whether pertinent or not. In the past, such articles had to be addressed individually. Now, a Philanthropy Data Analyst is presented with these clusters, accompanied by a probability score indicating the relevance of a given group. This ensures that the most pertinent clusters are prioritized and tackled first.

3 Future AI integration

With a notably enhanced workflow and heightened productivity, the logical progression is to ask whether the role of Philanthropy Data Analysts could be eliminated. While complete substitution using the current methodology is impossible, revising the methodology to fit the capabilities of the available AI technologies is worthwhile considering. Given the substantial expenses

associated with human labor, even in the Western Balkans' middle-income nations, relying solely on an AI-assisted approach may not be a viable long-term strategy. Two primary avenues exist for achieving full autonomy.

The first method entails a profound transformation of our current methodology to better align with the AI-driven capabilities at our disposal. This could involve various modifications, ranging from simplifying specific classifications to reconfiguring how we calculate the aggregate donation sum for a nation. For the elimination of the role of Philanthropy Data Analyst, the changes in methodology would need to be radical. Rethinking this role would probably be more realistic as no matter how the methodology changes, a human will always have to be in the loop.

The second strategy suggests supplanting Philanthropy Data Analysts with AI agents underpinned by services offering API access to generative large language models (LLMs) like ChatGPT. Fundamentally, some tasks within a Philanthropy Data Analyst's purview could be deconstructed into discrete operations that a powerful model like ChatGPT-4 could effectively manage. Our preliminary experiments with ChatGPT-3.5 turbo—a cost-effective choice provided by OpenAI via its API—indicate its robust capacity. When provided with appropriate prompts, it can adeptly categorize donation instances. The rapid progress of these models, translating to cheaper and evermore capable agents, makes this a promising avenue for future implementation. However, although the AI agents could be performant and effective, the solution is far from substituting the tasks and domain expertise of Philanthropy Data Analysts.

In a practical scenario, achieving further automation would likely necessitate a hybrid of both approaches. This means radical methodology changes with human (domain expert) oversight.

4 Problems with AI

4.1 Cost of AI

The primary benefit is the potential reduction in costs. ChatGPT costs depend on the specific model used (whether ChatGPT-3.5 turbo or ChatGPT-4) and the context window, ranging from 4,000 to 128,000 tokens. The price fluctuates between $0.002 and $0.03 per 1,000 tokens. Thus, classifying donation instances using OpenAI's API could cost anywhere from a few cents to $3.84 for each donation instance. This calculation assumes that media articles related to a particular donation have been collated and deduplicated before classification. To illustrate, Serbia had 4,557 recorded donation instances. If all these were processed using the most advanced OpenAI model with the largest context window, the cost for just this one country would be approximately $17,498.88 annually. Compared with the average gross salary in Serbia, which stands at $12,873.72—with a considerably lower median figure (Statistical Office of the Republic of Serbia, 2023)—this method appears less cost-effective. However, opting for the less advanced OpenAI GPT-3.5 Turbo model with a 16,000-token context window would incur a cost of only $218.7, presenting a far more economical alternative.

Utilizing a standalone service from one of the open-source LLMs based on Llama 2 (an open model released by Meta) or its derivatives could be considerably costly, especially considering the cloud resources required by its 70-billion-parameter version or even more cost-effective options (such as quantized variant such as Vicuna with 13 billion parameters). This does not even account for the costs of building, maintaining, and upgrading the system. While there are hosted services that offer API access, their charges are often on par with, if not exceeding, those of OpenAI's GPT-3.5 Turbo. OpenAI has recently released GPT-4 Turbo, which has a larger context size and, more importantly, is cheaper than GPT-4. The question of cost and whether API access to

proprietary or self-hosted open-source models is more affordable depends on each option's capabilities. The open-source alternative could be cheaper if smaller models show a similar level of capability as the proprietary models. For the time being we have not tested any of the open-source models, but benchmarking several different models for our use case against ChatGPT would be straightforward. Given the rapid advancement and the size of the community gathered around these open-source models, it will likely be a cost-effective and performant option in the future if it is not already.

Equally significant are the human resources expenses tied to the development of these infrastructures. While Catalyst Balkans crafted its system in-house, this endeavor redirected resources from other essential programs and services. Most non-profits typically lack the in-house capabilities even to construct their websites, much less having data scientists and engineers readily available. Assembling a team to create—and continuously develop and maintain—such a system could entail hefty expenditures.

4.2 Rapid innovation and changes to LLMs

Pipelines are susceptible to disruptions when models change. Even if these modifications generally enhance the models, each update necessitates a reassessment of the AI pipeline in place, which might need adjustments to align with the revised model. OpenAI has rolled out "frozen" models, which are updated less frequently compared to their regular counterparts. Nevertheless, even these are not guaranteed to have indefinite availability.

4.3 Integration of the knowledge

The role of Philanthropy Data Analysts at Catalyst Balkans extends beyond merely processing media articles, coding donation instances, and verifying them. Over time, these analysts have evolved into experts on charitable giving within their respective countries. While we provide data access as a service to our ecosystem, we frequently encounter inquiries that only someone deeply versed in the nuances of giving within a specific country can address.

Furthermore, these analysts have cultivated an in-depth understanding of entities in their designated countries. Their knowledge transcends mere reportage, equipping them to piece together narratives and make informed decisions based on often scant media information. This depth of understanding is further enriched by their connections within these countries, particularly their communication with individuals affiliated with donor and beneficiary institutions.

5 Insights and implications

The shift from the original data collection process to an AI-assisted approach at Catalyst Balkans marks a significant evolution in our handling of philanthropic data. This transformation exemplifies the transition from traditional methods to technologically advanced techniques, highlighting the efficiencies and effectiveness of AI integration. In comparison, our new process helped us manage the problem of false positives. AI's precision in filtering irrelevant data drastically reduced the burden of sifting through irrelevant content, a significant challenge in the original method. Additionally, our implementation for each language enabled streamlining the data collection process across different linguistic contexts, overcoming a major hurdle of the traditional approach. Most notably, the AI-assisted method markedly improved time efficiency and resource utilization,

addressing the labor-intensive nature of manual data processing. Beyond operational efficiency, AI integration has led to the discovery of new insights and strategic possibilities in philanthropic data collection.

Moreover, the efficiency gained using the AI-assisted approach that we have developed allows for handling the increasing data volumes without a corresponding increase in resources. This case study serves as a guide for other non-profits contemplating AI integration. Embracing AI can effectively address longstanding challenges in data management and processing. While initially resource-intensive, custom AI solutions can offer substantial long-term benefits in terms of efficiency. Maximizing AI benefits involves using it as an enhancement to, rather than a replacement for, human expertise, at least at this moment, which might not hold if AI continues to progress at an ever-increasing pace. Catalyst Balkans' experience in integrating AI can offer some insights into the digital transformation of the philanthropic sector globally, encouraging a shift toward more data-informed strategies in philanthropy and demonstrating how digitalization, through AI, can optimize operations, reduce costs, and amplify impact in the philanthropic sector worldwide. This global perspective underscores the potential of AI not just in the Western Balkans but in the broader philanthropic landscape.

6 Conclusion

The pursuit of leveraging AI to assist in the philanthropy data collection process at Catalyst Balkans showcases the synergistic potential between cutting-edge technology and traditional data management practices. The Giving Balkans database embodies the synthesis of manual efforts with technological solutions. This integration not only streamlines the data collection process but also enhances the quality and comprehensiveness of the data, revealing a broader and more nuanced perspective of the philanthropic landscape in the Western Balkans. Our piecemeal strategy of incorporating AI components allowed us to retain human oversight while still benefiting from automation. This hybrid model provided a balanced solution that addressed resource constraints, ensured data accuracy, and facilitated scalability. Even as we have made significant strides in optimizing the data collection process, the landscape of AI remains dynamic. The ever-evolving capabilities of large language models and declining costs present exciting prospects for future iterations of the Giving Balkans data collection process. The journey was not devoid of hurdles. From managing false positives to handling data veracity, the imperfections of AI compelled us to refine our methodologies constantly. Moreover, the financial implications and rapid technological changes associated with AI underline the need for organizations to remain agile and forward-thinking. As we move closer to achieving a completely autonomous process, the ethical implications of such a system cannot be understated. The balance between automation and human judgment will ensure that the Giving Balkans database remains a reliable and unbiased resource. The successes and challenges encountered in this case study hold broader implications for the global intersection of AI and philanthropy. As technologies become more accessible, there is an opportunity for organizations worldwide to harness them to capture philanthropic trends, understand donor behaviors, and ultimately foster a culture of giving. In conclusion, the journey of integrating AI into the data collection processes at Catalyst Balkans is emblematic of the broader narrative of technological transformation in the non-profit sector. As we navigate the complexities and promises of AI, it becomes evident that the convergence of human expertise and machine intelligence can pave the way for a more informed, efficient, and impactful future in philanthropy.

Notes

1 Serbia, Croatia, Bosnia and Herzegovina, Albania, Macedonia, Kosovo, and Montenegro.
2 https://donacije.rs is a crowdfunding website run by Catalyst Balkans, where non-profits can put up their campaigns and crowdfund for specific projects. Catalyst Balkans guides non-profit campaign owners throughout the process—helping them formulate the campaign, create content, and teach them how to fundraise.

References

Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics*. https://doi.org/10.48550/ARXIV.2008.09470

Bajra, U. Q. (2021). The interactive effects of remittances on economic growth and inequality in Western Balkan countries. *Journal of Business Economics and Management*, *22*(3), 757–775. https://doi.org/10.3846/jbem.2021.14587

Brinkerhoff, J. M. (2014). Diaspora philanthropy: Lessons from a demographic analysis of the Coptic diaspora. *Non-Profit and Voluntary Sector Quarterly*, *43*(6), 969–992. https://doi.org/10.1177/0899764013488835

Catalyst Balkans (2020, March 25). *Donacije.rs—COVID-19*. https://www.donacije.rs/covid19/

Catalyst Balkans (2021, November 30). *Giving Balkans: Philanthropy's Response to COVID-19 (September 30, 2021)*. Giving Balkans. https://givingbalkans.org/content/giving-balkans-philanthropy%E2%80%99s-response-covid-19-september-30-2021

Catalyst Balkans (2023). *Giving Balkans Database on Philanthropy in the Western Balkans.* https://givingbalkans.org/

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27. https://doi.org/10.1145/1961189.1961199

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2023). *Shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny

Drobarov, R., Popovska, B., & Gelev, I. (2021). Impact of COVID-19 on sustainability of civil society organizations in the Western Balkan Region. *Bezbednost, Beograd*, *63*(3), 57–76. https://doi.org/10.5937/bezbednost2103057D

Duncan, D. (2016). Language policy, ethnic conflict, and conflict resolution: Albanian in the former Yugoslavia. *Language Policy*, *15*(4), 453–474. https://doi.org/10.1007/s10993-015-9380-0

EWB. (2023, May 24). Freedom house: Democratic institutions in the Western Balkans continued to falter in 2022. *European Western Balkans*. https://europeanwesternbalkans.com/2023/05/24/freedom-house-democratic-institutions-in-the-western-balkans-continued-to-falter-in-2022/

Freedom House. (2016). *Back Where We Started in the Balkans*. Freedom House. https://freedomhouse.org/article/back-where-we-started-balkans

Galjak, M. (2020). Dva primera upotrebe teorije grafova u društvenim naukama: Analiza interakcija na Tviteru tokom izbora 2016. U Srbiji i analiza GivingBalkans podataka o filantropiji na Zapadnom Balkanu. In V. Mentus & I. Arsić (Eds.), *Promišljanja aktuelnih društvenih izazova: Regionalni i globalni kontekst* (pp. 232–251). Institut društvenih nauka.

Glenny, M. (2001). *The Balkans: Nationalism, War, and the Great Powers, 1804–1999*. Penguin Books, New York.

Grødeland, Å. B. (2006). Public perceptions of non-governmental organizations in Serbia, Bosnia & Herzegovina, and Macedonia. *Communist and Post-Communist Studies*, *39*(2), 221–246. https://doi.org/10.1016/j.postcomstud.2006.03.002

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). *FastText.zip: Compressing Text Classification Models*. https://doi.org/10.48550/ARXIV.1612.03651

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). *Bag of Tricks for Efficient Text Classification*. https://doi.org/10.48550/ARXIV.1607.01759

Lau, J. H., & Baldwin, T. (2016). *An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation*. https://doi.org/10.48550/ARXIV.1607.05368

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *CoRR*, *abs/1405.4053*. https://doi.org/10.48550/arXiv.1405.4053

Marinković, I., & Galjak, M. (2021). Excess mortality in Europe and Serbia during the COVID-19 pandemic in 2020. *Stanovništvo*, *59*(1). https://doi.org/10.2298/STNV2101061M

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. https://doi.org/10.48550/ARXIV.1802.03426

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. https://doi.org/10.48550/ARXIV.1301.3781

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, *10,* 61–74.

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2), 2.

Statistical Office of the Republic of Serbia (2023, August 25). *Average Salaries and Wages Per Employee, June 2023*. https://web.archive.org/save/https://www.stat.gov.rs/en-us/vesti/statisticalrelease/?p=13675&a=24&s=2403?s=2403

Transparency International (2020, December 11). *Captured States in the Western Balkans and Turkey—News*. Transparency.Org. https://www.transparency.org/en/news/captured-states-western-balkans-turkey

USAID. (2023). *Civil Society Organization Sustainability Index (Reports)*. FHI 360. https://www.fhi360.org/sites/default/files/media/documents/csosi-europe-eurasia-2021-report.pdf

World Bank (2023). *GDP Per Capita (Current US$)—Kosovo, Serbia, Croatia, North Macedonia, Albania, Montenegro, Bosnia and Herzegovina* [dataset]. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=XK-RS-HR-MK-AL-ME-BA