Unveiling Communication and Support Dynamics: Analyzing Telegram and Helpline Data in Conflict Zones

A Major Qualifying Project submitted to:

The Faculty of Worcester Polytechnic Institute

in Partial Fulfillment of the

Bachelor of Science Degree

Project Advisor:

Professor Renata Konrad

Author:

Atharva Tiwari

Abstract

The traditional Fordist approach to humanitarian response often falls short in addressing the complexities of modern crises. To better understand the demands arising from the Ukrainian situation and suggest practical solutions, this project leverages the Telegram and 527 datasets. By conducting rigorous analysis, we aim to offer insights through the following approaches. (1) 527 Helpline Dataset Analysis: We employed Exploratory Data Analysis (EDA) and a regression model to understand the factors influencing the protection of Ukrainian rights abroad. The regression model helping identified key features impacting these protections. (2) Telegram Dataset Analysis: Using EDA and a Latent Dirichlet Allocation (LDA) machine learning model, we analyzed the Telegram dataset to uncover important themes and communication patterns. This analysis sheds lights on public debates and concerns. By integrating the findings from both datasets, this project aims to provide a comprehensive understanding of digital communication trends and practical support mechanisms. These insights have significant implications for future research on conflict-affected populations and international human rights efforts, offering a nuanced perspective on addressing modern humanitarian crises.

Executive Summary

This project aims to provide valuable insights into communication patterns and assistance mechanisms for populations affected by military conflict. By investigating two interconnected datasets – Telegram and the 527 dataset– we explore urgent real-world issues with a comprehensive approach. The Telegram and 527 hotline datasets were selected for this project for their relevance and richness in shedding light on critical aspects of the Russian full-scale invasion of Ukraine. The Telegram dataset, consisting of communications from various users, offers a unique opportunity to understand public sentiment, trends, and behaviors on a widely-used social media platform. This dataset was analyzed using Exploratory Data Analysis (EDA) and a Latent Dirichlet Allocation (LDA) machine learning model. The analysis of the Telegram dataset uncovers pivotal topics and communication patterns, illustrating public discourse and societal concerns surrounding the conflict in Ukraine.

The 527 hotline dataset, offered detailed records of helpline interactions, proving invaluable for analyzing the needs and concerns of individuals seeking assistance. This dataset provided crucial insights into factors influencing the protection of Ukrainian rights abroad, identified through feature importance in the regression model. It played a pivotal role in evaluating the effectiveness of helpline services, understanding caller demographics, and recommendations of how to provide address their needs. Collectively, these analyses provide a comprehensive understanding of digital communication trends and practical support systems, enabling a holistic approach to addressing the needs and challenges of the Ukrainian people during this full-scale invasion.

The Telegram dataset aimed to uncover recurring themes and patterns in public conversation, crucial for pinpointing variables affecting the defense of Ukrainian rights abroad. Analyzing the Telegram dataset during a time of geopolitical tension aimed to understand what people type of aid people were requesting and offering. Through EDA and LDA model, significant trends and themes emerged, shedding lights on key issues, popular concerns, and attitudes toward the situation in Ukraine. This analysis offers valuable insights for scholars and policymakers interested in digital communication during crises and how traditional humanitarian operations use these insights to adapt to a rapidly changing conflict and address the needs of people. Similarly, the 527 helpline dataset study focused on identifying variables influencing the defense of Ukrainian rights abroad. The EDA revealed important trends, and a regression model quantified the significance of various features, highlighting key factors such as discussion nature and individuals' legal status (e.g., refugee status),. Such findings are important for organizations focusing on human rights and

support services, as it identifies potential areas of greatest effectiveness for interventions.

By integrating data from both datasets, this study provides a comprehensive understanding of the interplay between support mechanisms and public communication. The 527 hotline offers a micro-level view of individual support needs and responses, while the Telegram dataset provides a macro-level insight of public attitude and discourse. Together, these evaluations underscore the importance of coordinated efforts in digital communication and local support services. This project also opens avenues for future research. Tracking changes in themes within the Telegram dataset over time may shed light on shifting public concerns. Furthermore, further investigation into specific interventions offered by the 527 helpline may uncover the most effective methods for assisting individuals affected by the conflict. Given current geopolitical tensions, similar approaches could be applied to understand communication patterns, support requirements and needs for certain services in other contexts, such as the Israel- Palestine conflict.

In summary, this MQP demonstrates the efficacy of combining machine learning with exploratory data analysis to derive relevant meaningful insights from diverse datasets. Policymakers, human rights organizations, and scholars dedicated to understanding and assisting individuals in crisis zones will find these findings particularly pertinent

Acknowledgements

This project would not have been possible without the guidance, support and encouragement from many peoples. First, I would like to thank Renata Konrad, Ph.D. (Worcester Polytechnic Institute) and Laura Dean, Ph.D. (Millikin University) for their valuable feedback, kindness and guidance throughout the project. I would also like to thank Amir Jamali (WPI, Doctoral Student, School of Business) and Solomiya Sorokotyaha (WPI, Master Student, School of Business) for their work on preprocessing the Telegram data as well as their support to the project and feedback.

Contents

List of Figures

1 Introduction

1.1 Understanding Refugees and Volunteer Response

The United Nations defines a refugee as "someone who has been forced to flee his or her country because of persecution, war or violence. (UNHCR, 2017)." Most of the time, individuals flee their country due to war, tribal and religious violence, and political revolutions, to name the major causes. Furthermore, the United Nations reports that 52% of the world's refugees come from Syria, Ukraine, and Afghanistan (UNHCR, 2017). Since 1975, more than 3 million refugees have been resettled throughout 50 states in the United States. (UNHCR, 2020) It is also important to note that this displacement makes this population vulnerable to trafficking situations. This means that these individuals could be transported by the use of force, fraud, or coercion to obtain some work. According to the United Nations Basic Needs Approach, refugees and internally displaced persons require essential services and resources such as health, nutrition, WASH (water sanitation, and hygiene), shelter, energy and domestic items when they are forced to leave their homes (UNHCR, 2018). With refugees being displaced for an average of 10-26 years, there are critical long-term needs that must be addressed. These include integrating refugees into the local community, and securing steady employment jobs to provide a stable source of income to prevent them from falling into poverty (Ferris, 2018). In an era dominated by mass media, the stories, journeys, and experiences of these individuals often go unnoticed by society at large. Globally, and particularly in the West, the concept of refugees elicits diverse reactions. Studies conducted by the Pew Research Center highlight this divide. In Western countries, public opinion is split on whether immigrants contribute positively to their country. In the US, 34% believe immigrants are a burden and 59% believe that they are not (Lipka, 2022). Many people associate refuges with concerns about crime and terrorism. In Europe, 59% of individuals believe that refugees increase the risk of terrorism, while 36% believe that they do not (Poushter, 2016). While some individuals may be reluctant to support refugees, volunteers often rise to the occasion when a disaster occurs, becoming the first ones to help. These volunteer efforts play a crucial role in providing immediate aid until established aid organizations such as the UN, and Red Cross can step in.

During times of disaster, volunteers are often among the first to respond. For example, during the COVID pandemic, volunteers stepped up to assist overburdened/over-capacity hospitals and provide financial support for families. A study in the United Kingdom found that during the pandemic 12.4 million individuals volunteered to help in these situations (ICRC, 2022). In Ukraine, volunteers played a crucial role in the initial weeks of the war, significantly contributing to the humanitarian effort. Due to this volunteer effort, a large humanitarian catastrophe was circumvented (Dunn and Kaliszewska, 2023). Another

example is the recent Israel-Hamas war. Following the outbreak of the attacks, thousands of grassroots volunteer initiatives appeared across the country to help those in need (Kershner and Shaar-Yashuv, 2023). Some of these efforts include distributing clothes and medicine, bulk-producing meals, collecting resources (chargers and clothes) for soldiers, and providing psychological support for victims (Kershner and Shaar-Yashuv, 2023). While such volunteer efforts are heartwarming, the question arises: where are the established aid organizations? Groups such as the UN, Red Cross, and Doctors without Borders are known for their large-scale humanitarian aid efforts. however, in violent conflicts such as those in Ukraine and Gaza, the chaotic, dangerous environment can create operational challenges for these organizations, making it difficult to operate effectively. (ICRC, 2022). In Ukraine, large formalized aid was unable to be delivered to internally displaced and evacuated individuals on a large scale until humanitarian corridors could be established (Amnesty International, 2022). As the bloody Israel-Hamas conflict continues, the people of Gaza, are facing severe shortages of essential resources such as water, food, and medical supplies, due to the Israeli blockade. Furthermore, the infrastructure of Gaza is on the nearing collapse with minimal electricity avialble. According to medical teams in Gaza, the medical system is on the brink of failure. (Haq, 2023) Due to this urgent need for aid, humanitarian groups are calling on both parties in this conflict to establish concrete humanitarian corridors for aid to be delivered (Haq, 2023). Recent conflicts and disasters have shown that volunteer support is crucial in providing urgent aid until large aid organizations can arrive. One of the causes of the delayed involvement of aid organizations is the lack of urgency from parties involved in the conflict to create these humanitarian corridors which help provide large amounts of aid to impacted populations. Next, we will discuss the Ukrainian refugee crisis and its causes and effects.

1.2 Ukrainian Context

In February of 2022, Russia significantly escalated its ongoing conflict with Ukraine, which began in 2014 by laughing a full-scale invasion of the country. Since then, fighting has spread to various parts of the country including major cities like Kharkiv, Mariupol, and Kyiv. This has resulted in extensive property and infrastructure damage, such as the destruction of the Kakhovka Dam, leading to mass flooding in the city of Kherson. Furthermore, 50% of the electrical grid was damaged leaving lots of citizens with no power. Additionally, many people have died and been displaced, both internally (to somewhere else in Ukraine) and externally (to other countries). As of May 2023, those who are displaced internally in Ukraine number 5.1 million (UNHCR, 2023), while those who are refugees (externally displaced) number 6.2 million in June 2023 (UNHCR, 2023). In total, about a quarter of the country's 43 million people are displaced.

After the initial weeks of the 2022 full-scale invasion, there were 5.1 million internally displaced

persons and 6.2 million refugees. Internally, Ukrainians were trying to flee the frontlines of the Russian advance. Externally, Ukrainians were trying to escape the country to avoid war and move far away from the conflict. However, with this surge in refugee activity at Medyka, one of the busiest border crossings, there was no response from the United Nations High Commission for Refugees (Dunn and Kaliszewska, 2023). For example, when Dunn and Kaliszewska (2023) visited multiple border crossings in Poland, they observed that large aid organizations and central governments were absent from providing aid to displaced individuals. They assert that "humanitarian catastrophe" was avoided due to the significant volunteer response they witnessed at the border (Dunn and Kaliszewska, 2023). Other sources including Amnesty International, the New York Times and Refugees International discuss the border crisis and how the first-hand volunteer response in Ukraine is much more effective than that of large centralized aid organizations.

1.3 Understanding of Fordist Humanitarian Organizations

But why was the volunteer response more effective? To answer this, it is important to understand the nature of large humanitarian aid organizations and how they operate. A core principle of these types of humanitarian organizations is the concept of Fordism. Fordism gained popularity after World War II exemplified by the Ford Motor Company's use of mass production to achieve economies of scale (Jessop, 2020). Since most modern humanitarian agencies were founded in this post-war era, they adopted Fordist practices that remain today.

These Fordist-era humanitarian organizations operate similarly to businesses but differ in that they must respond almost immediately during times of emergency. This operational mandate requires them to "preposition large warehouses of standardized goods near likely conflict and disaster zones. However, this centralized and standardized approach can sometimes hinder their responsiveness compared to the more immediate and flexible efforts of grassroots volunteers (Dunn and Kaliszewska, 2023). Due to these reasons it can be deduced that "humanitarian aid is 80% logistics and it can be achieved through efficient and effective operations"; this is one of the reasons why refugee camps developed (Van Wassenhove, 2006, para.2). These refugee camps served as a perfect springboard for the transport and distribution of humanitarian aid. One of the reasons for the inefficiency of these large humanitarian organizations can be attributed to the rigidity of Fordism. At the turn of the century, the average length of displacement wavered around 26 years. As a result, displaced persons avoided refugee camps and moved into cities where they could find better job and educational opportunities (UNHCR, 2017). Additionally, many refugees in urban areas were able to meet their needs through markets in cities, assistance from churches, and other local organizations. This was exactly the case in Ukraine. In their paper, Dunn and Kaliszewska (YEAR) cite a Facebook post in which

a Polish volunteer says:

> The war in Ukraine also proved their inability to act fast and do anything on a large scale. They are still in the process of "assessing the needs" in Ukraine, just like they were when thousands of "non-professionals" rushed to help. It didn't stop (them) from fundraising for Ukraine, and effectively diverting the funds from smaller groups that did something real (Dunn and Kaliszewska, 2023).

This highlights the issue that many Fordist organizations are having in the 21st century, to transition to a post-Fordist model for providing aid.

1.4 Effectiveness of Volunteer Response

A key factor in the robust volunteer response to the crises has been the role of social media. Since the onset of the full-scale invasion, the ensuing humanitarian crisis has been extensively covered on various social media platforms. Both small victories and losses for each side are discussed not only in mainstream media as well as smaller social media channels including Facebook groups, subreddits, Telegram chats and Discord servers.

These platforms have become essential resources for families and individuals fleeing their homes, providing information on available assistance and connecting them with people and organizations willing to help. A recent study from the University of Copenhagen in Denmark highlights that during the first few weeks of the war many support groups in Denmark emerged to organize support for Ukrainian refugees (Hjalmar Bang Carlsen and Toubøl, 2023). Based on a more in-depth analysis Carlsen's team found that these groups were merely performative; individuals genuinely were inclined to help Ukrainians resulting in clear and concrete actions (Hjalmar Bang Carlsen and Toubøl, 2023). Another study conducted by a team at the University of Southern California found that social media can be a powerful tool for grassroots efforts during times of humanitarian distress. Their research details how activists wanting to help Ukrainians use platforms such as Facebook, Twitter, Instagram and YouTube. These sites generate significant traffic from one site to another primarily to raise awareness to different aspects of the humanitarian effort. For example, users are directed to sites like Patreon, Go Fund Me and Google Forms encouraging individuals to sign up as well as donate to help with the humanitarian effort (Ye et al., 2023).

One key factor in making the volunteer response was the age of social media and the role that they played in helping refugees in this crisis. Since the beginning of the war, the Ukrainian-Russian conflict, including the ongoing humanitarian crisis, has been heavily covered on social media. Small victories and losses for both sides are talked about in mainstream media as well as "smaller" social media channels. These

channels include but are not limited to Facebook groups, subreddits, Telegram chats as well as Discord servers. Alongside discussions of the war, social media has been a place for families and individuals escaping their city, village, or even the country to look for people and organizations that will help their families and offer to help them. A recent study at the University of Copenhagen in Denmark claims that during the first few weeks of the war, many support groups in Denmark emerged to organize support for Ukrainian refugees (Hjalmar Bang Carlsen and Toubøl, 2023). Based on a more in-depth analysis Carlsen's team claims that these groups were not for show and individuals were inclined to help Ukranians with there being clear concrete action (Hjalmar Bang Carlsen and Toubøl, 2023). Another study performed by a team at USC found that social media can be a powerful tool for grassroots efforts during times of humanitarian distress. In their study, the team details how the main channels used by activists wanting to help Ukrainians are Facebook, Twitter, Instagram, and Youtube. Amongst these sites, there is traffic from one site to another primarily to raise awareness of different aspects of the humanitarian effort. For Example, from these sites there is traffic generated to other sites like Pateron, GoFundMe and Google Forms to encourage individuals to sign up as well as donate to help with the humanitarian effort (Ye et al., 2023).

Social Media networks are often interpreted only as being a part of major social media platforms like Facebook, X, and Instagram. However, social media networks often involve an individual's network as well. The use of an individual's network was another powerful tool that was used by volunteers in the Ukrainian refugee crisis to be able to assist those in need. One example of this is Dale Smith who owns a Kyiv-based oil-trading company, as soon as the war broke out he helped his employees flee the country (Dunn and Kaliszewska, 2023). A few days later he had bought all the stock from a medical warehouse in Poland and sent it to the border to help Ukrainian refugees leaving the country. Aleksandra Lewandowska was a Polish trader who worked for Smith and she had helped arrange for supplies to be purchased and sent to Ukraine using her own social network in Poland. Another employee of Smith's used their network to find an official who works for the Ukrainian border patrol to ensure that the goods made it over the border. Dale Smith was able to successfully send over $2 million USD of aid in the first few weeks of the war, this was during a time when UNHCR had not even responded (Dunn and Kaliszewska, 2023). Dunn and Kaliszewska claim that many others like Dale Smith were able to successfully help Ukrainians whereas large humanitarian organizations were unable to do so. In sum, the reason why the volunteer response through social networks (both digital and personal) was more effective than that of large humanitarian was because "(1) They moved a high volume of aid fast; (2) They delivered aid in response to orders placed by people in need themselves; (3) they were based on trust, not accountability (Dunn and Kaliszewska, 2023)."

This section explained what refugees are and why they flee their country. Additionally, this section dove deeper into the subject of our study, understanding and analyzing volunteer networks, and understand-

ing how these volunteer networks function to recommend how established aid organizations can improve their response. The problem that our project seeks to address is below.

1.5 Problem Statement and Objectives

Problem Statement: The problem is that the Fordist way of looking at humanitarian response doesn't always work. Using the Ukrainian situation, the project's goal is to identify and use the Telegram and 527 dataset to better understand what is needed, and to propose appropriate strategies to address those needs.

Objective 1: Examine social networks on social media (Telegram).

Objective 2: Characterize hotline statistics, how did the full-scale invasion impact the need for information and help for Migrants and Internally Displaced Persons across Ukraine.

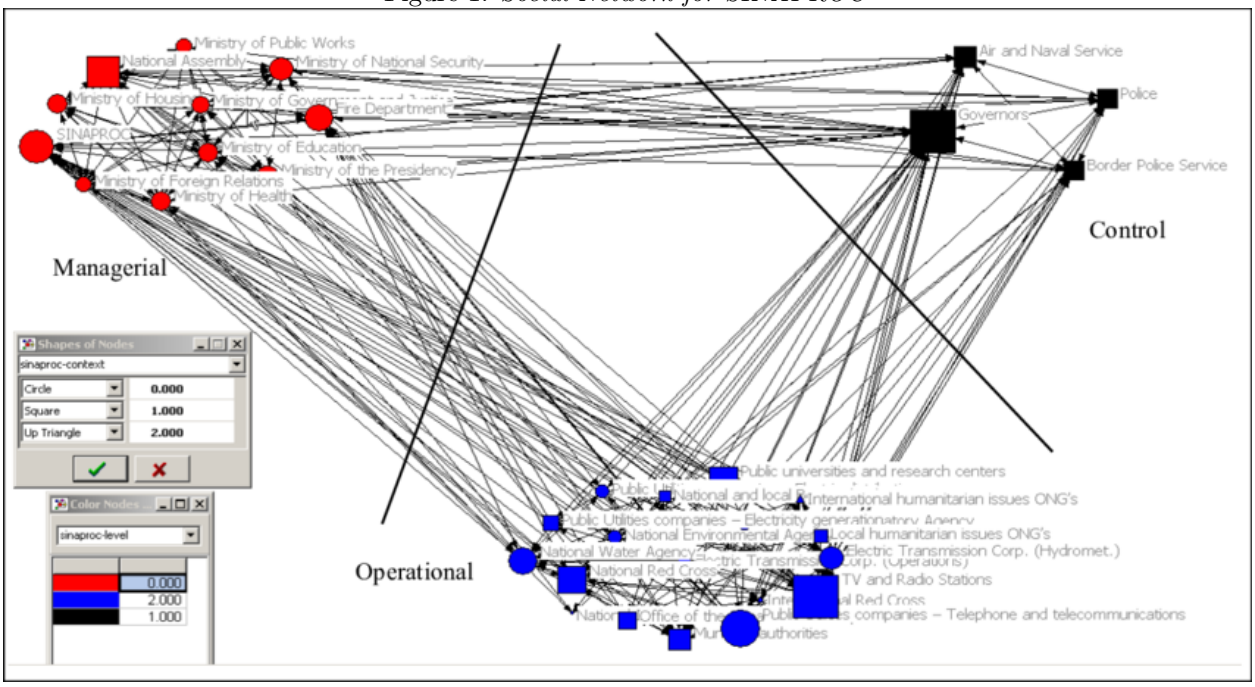
Objective 3: Provide recommendations for how the analyses can be interpreted to better serve vulnerable communities.

2 Related Work

2.1 Social Network Analysis on Humanitarian Operations

Several studies have utilized Social Network Analysis (SNA) to examine humanitarian networks. One such study in 2013 applied SNA to analyze the communication patterns involved in humanitarian logistics operations in Latin America (Alvarez and Serrato, 2013). In this study, the research team detailed their methodology for generating graphs to compare and contrast the different network structures among humanitarian logistics organizations in different countries. The team focused on organizations at the managerial, control, and operational levels within specific countries, developing a multilevel communication model. This model, combined with SNA was designed to illustrate both the intra-level communication and the inter-level communications during various stages of the humanitarian relief process. To assess the efficiency and effectiveness of a particular network Alvarez et al. used common measures such as centrality, closeness, and betweenness among organizations. In their paper the authors highlight an example of running their methodology on data that they obtained from Panama, the resulting social network diagram is shown in Figure 1 below.

Figure 1: *Social Network for SINAPROC*



Note. From "Social Network Analysis for Humanitarian Logistics Operations in Latin America" by H. Alvarez and M. Serrato (2013). IIE Annual Conference and Expo 2013, https://www.researchgate.net/publication

In their resulting analysis of this data, the team noticed that "governors, telephone companies, and TV and Radio Stations have the greatest level of out-degrees, while TV and Radio Stations also have, by large, the greatest level of in-degrees", based on their definition this meant that these "actors" were influential players in the spreading of information with regards to humanitarian logistics (Alvarez and Serrato, 2013). Based on this data, the team asserts that in Panama the logistics of humanitarian relief are highly dependent on the communication networks. Despite the contributions of this study, it is important to recognize there are limitation. One signifcant drawback is their analyses was confined to a single phases of the humanitarian relief cycle. In their paper, they acknowledge that an area for future study would be to apply their methodology to different stages of the cycle to see which organizations are key in humanitarian response as well as how governments can involve smaller organizations (Alvarez and Serrato, 2013). Overall, this study showed how Social Network Analysis can be effective in determining the potency and efficiency of humanitarian networks.

A recent study at USC employed SNA to examine online support networks. Using a data set of 68 million posts, the researchers explored the connections not only within social media networks but also among crowd sourcing, and crowdfunding websites Ye et al. (2023). To collect the data the team parsed through data using respective proprietary APIs for major social media platforms like X, Facebook, Instagram, and YouTube, this includes any embedded links. These embedded links were key to this study as this enabled researchers to detect which sites were the most common external links apart from links between previously listed major social media platforms. From Figure 2 below one can see that across the major social media platforms there are over 13.9 million embedded URLs.

Figure 2: *Number of posts, unique URLs, and web domains shared in every social media under analysis.*

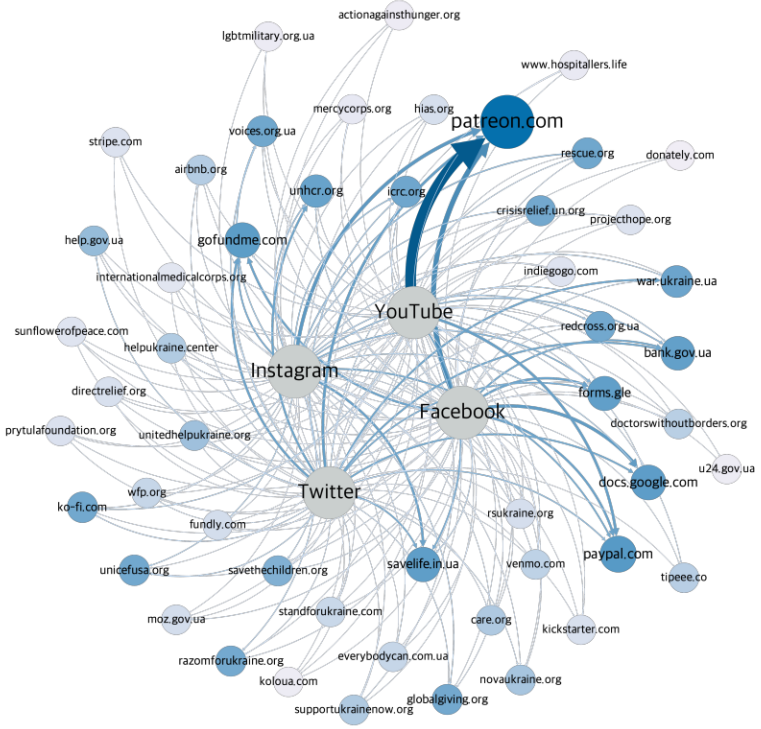
Platform	Posts	URLs	Domains
Twitter	55,938,686	8,154,714	285,165
Facebook	11,777,025	5,542,833	185,567
Instagram	68,403	37,655	18,103
Youtube	275,937	204,344	30,405

Note. From "Online Networks of Support in Distressed Environments: Solidarity and Mobilization during the Russian Invasion of Ukraine" by J. Ye et al (2023). University of Southern California, https://doi.org/10.48550/arXiv.2304.04327

To refine the volume of links, the team only considered links "recommended by highly credible media outlets as well as information aggregation web pages from credible sources. (Ye et al., 2023)." Furthermore, to further reduce the pool of data to more reliable sources the team only considered the support sites that

"appeared on at least two of the four major platforms (Ye et al., 2023)." To further break down support sites, based on observational analyses they were classified based on their fundraising mechanics. The different categories were: crowd-sourcing platforms (Google Docs, Google Forms, Cryptocurrency Donations), crowdfunding platforms (GoFundMe, Patreon), government fundraising sites (National Bank of Ukraine), Local NGOs and Funds, and International NGOs and funds (Hjalmar Bang Carlsen and Toubøl, 2023). Further analysis of this data indicated that Twitter had the greatest number of shared URLs with 77.10%, Facebook with 18.76%, and YouTube and Instagram with 2.51% and 1.62% respectively (Ye et al., 2023). To supplement these findings Alvarez et al., constructed a graph of the major social media networks and their inter-connectivity. This graph is seen in Figure 3 below.

Figure 3: *The online network of support: Connections between Social media platforms and support sites are represented with normalized, directed, and weighted edges.*



Note. From "Online Networks of Support in Distressed Environments: Solidarity and Mobilization during the Russian Invasion of Ukraine" by J. Ye et al (2023). University of Southern California, https://doi.org/10.48550/arXiv.2304.04327

In the graph above, the darker and larger a node,the more incoming links the support site has. The four gray nodes represent the source social media platforms. The study discovered that Twitter (X) and Facebook had an equal distribution of crowdfunding and crowd-sourcing sites, whereas Instagram and YouTube have more links to crowdfunding sites. The researchers attribute this finding to the visual nature of Instagram and

YouTube. To further understand the connections between major social media platforms, the team performed a Granger-causality test on the "volume of posts linking to different support sites shared on the four networks (Ye et al., 2023)." The Granger-causality test indicated that Twitter Granger causes the greatest number of support sites to the destination platforms. It is important to note that type of support site (crowdfunding, crowd-sourcing, and other sites) influences the Granger-causality outcomes. Figure 4 illustrates the results of the Granger-causality test, highlighting these variations.

Figure 4: *Pairwise Granger causality results between platforms. The maximum lag parameter is set to 3 and the significance level to p ¡ .05 (Note: p ¡ .05, p ¡ .01, p ¡ .001). # = Total number of support sites that the source platform Granger-causes the destination platform,CS = Crowd-sourcing Sites (Google Docs and Forms), CP = Crowdfunding Platforms, OCS = Other Crowdfunding Sites (Government, Local and International NGOs/Funds). # = CS + CP + OCS.*

Source	Destination	#	Domains	CS	CP	OCS
Twitter	Facebook	6	fundly.com**, rsukraine.org*, globalgiving.org**, mercycorps.org*, novaukraine.org**, moz.gov.ua*	0	1	5
Facebook	Twitter	6	docs.google.com*, rescue.org**, unhcr.org*, redcross.org.ua**, directrelief.org*, unitedhelpukraine.org*	1	0	5
Twitter	Instagram	7	docs.google.com*, forms.gle***, patreon.com*, globalgiving.org*, mercycorps.org*, moz.gov.ua**, doctorswithoutborders.org**	2	1	4
Instagram	Twitter	6	voices.org.ua**, care.org***, icrc.org***, directrelief.org*, airbnb.org*, sunflowerofpeace.com*	0	0	6
Twitter	YouTube	14	docs.google.com*, patreon.com**, venmo.com*, tipeee.co*, voices.org.ua***, rsukraine.org***, globalgiving.org***, rescue.org*, prytulafoundation.org*, crisisrelief.un.org***, hospitallers.life***, novaukraine.org*, koloua.com*, standforukraine.com***	1	3	10
YouTube	Twitter	8	paypal.com**, ko-fi.com*, bank.gov.ua**, savethechildren.org*, doctorswithoutborders.org**, care.org*, unitedhelpukraine.org*, moz.gov.ua***	0	2	6
Facebook	Instagram	8	kickstarter.com**, venmo.com**, ko-fi.com*, helpukraine.center***, internationalmedicalcorps.org*, redcross.org.ua***, directrelief.org*, hias.org***	0	3	5
Instagram	Facebook	6	voices.org.ua***, rsukraine.org***, care.org*, moz.gov.ua**, actionagainsthunger.org*, sunflowerofpeace.com**	0	0	6
Facebook	YouTube	8	gofundme.com*, bank.gov.ua*, helpukraine.center*, unhcr.org*, standforukraine.com**, rsukraine.org***, directrelief.org*, koloua.com*	0	1	7
YouTube	Facebook	8	paypal.com*, venmo.com*, globalgiving.org***, savethechildren.org*, unicefusa.org*, internationalmedicalcorps.org**, hospitallers.life**, supportukrainenow.org**	0	2	6
Instagram	YouTube	12	savelife.in.ua***, everybodycan.com.ua***, supportukrainenow.org*, rsukraine.org***, care.org*, globalgiving.org**, icrc.org**, projecthope.org***, doctorswithoutborders.org*, redcross.org.ua*, novaukraine.org***, moz.gov.ua***	0	0	12
YouTube	Instagram	5	bank.gov.ua*, rescue.org***, wfp.org**, unhcr.org***, prytulafoundation.org***	0	0	5

Note. From "Online Networks of Support in Distressed Environments: Solidarity and Mobilization during the Russian Invasion of Ukraine" by J. Ye et al (2023). University of Southern California, https://doi.org/10.48550/arXiv.2304.04327

Further analysis of these support sites delves into a discussion of the advantages and disadvantages of crowd-sourcing vs. crowdfunding during disaster response. Their analysis indicated that crowdsourcing offered an organized approach to coordinating aid. Platforms such as Google Forms and Docs were utilized to direct users to sites to donate, sign petitions, help pets and families, or even register as a volunteer on the front lines. In contrast, crowdfunding platforms were employed by anonymous individuals, and smaller NGOs to raise funds for Ukrainian families and individuals who lacked sufficient resources or could not receive aid from official sources (Ye et al., 2023). This study is an excellent example of how SNA can be

utilized to understand volunteer networks and the role of social media in disaster response. In this case, SNA was used to understand how major social media platforms generated traffic to support sites to fund raise and gain volunteer support. Despite the success of this study one of the limitations that the authors state is that they did not consider Telegram channels which could have been a source of valuable data. Additionally, the authors suggest further research should investigate the spread of misinformation and false support for volunteer networks and compare those findings.

This section highlights which studies have been conducted regarding the use of SNA to understand volunteer and humanitarian networks during times of disaster. Furthermore, the study conducted by Ye et al., helps the reader understand the role of social media in volunteer efforts. Although these studies demonstrate a large repository of data being used, our study will use data obtained from Telegram. These studies prove how social media is used during crises; although this MQP won't be conducting an SNA, we recognize the importance this analysis plays.

2.2 Spontaneous Volunteer Response

Typically, when a disaster occurs, spontaneous volunteers arrive en-masse provide immediate assistance, while established aid organizations take longer to set up and organize thier efforts (Daddoust et al., 2021). This pattern has been observed in numerous past events, including the 2004 Indian Ocean Tsunami and the 2015 Nepal Earthquake (Daddoust et al., 2021). The majority of initial rescue operations were done with spontaneous volunteers. While these responses are crucial and effective, they can pose challenges when the number of volunteers is very high and there are few or no leaders to coordinate them. Traditional aid organizations tend to avoid working with spontaneous volunteers because they can be a liability. Issues such as untrained personnel, lack of leadership and dynamic conversation, and the overwhelming number of volunteers in a large disaster can hinder the efforts of trained professionals (Daddoust et al., 2021).

Daddoust et al. (2021) also conducted a questionnaire to understand why Ontario Emergency Responders were hesitant to allow spontaneous volunteers to help. They found that roles requiring a certain amount of skill can become liabilities if volunteers are not properly trained. Liabilities include potential injury of volunteers and breaches of legislation. The paper concludes that while spontaneous volunteers can be effective, and bring valuable knowledge, such as awareness of local community norms, they still pose a liability to aid providers. These concerns must be addressed before spontaneous volunteers can work seamlessly with established NGOs.

2.3 The Role of Helplines in providing valuable insights in the anti-trafficking space.

While helplines are not as widely studied as social media channels; they offer valuable insights into the minds and conditions of marginalized individuals or those in a difficult situation. Hotlines are crucial not only for anti- human trafficking efforts but also in everyday life, such as 988 (Suicide Prevention Hotline), Child Abuse and Neglect, and Domestic Violence. Such hotlines are typically affiliated with the state , managed by transnational organizations, or operated by NGOs. These hotlines are widely deployed around the world, and offer valuable insights into how marginalized populations are exploited by traffickers. More importantly, analytics derived from these hotlines can empower NGOs to better assest vulenable populations (Cockbain and Tompson, 2024). It is important to note that "helping" in this context involves educational campaigns aimed at raising awareness among vulnerable populations about potential pitfalls and how to avoid them. For example, the campaigns can include educational material on how to recognize suspicious adds and identifying authorities to contact in difficult situations. A recent study conducted by Cockbain and Thompson (2024) analyzed a dataset from a major anti-trafficking hotline. Thier research provided insights into what can be derived, propose future research in this space and how helplines can be important in a "complex system of anti-trafficking activities" (Cockbain and Tompson, 2024).

In the aforementioned study, the data from hotlines is well-monitored and has significant quality control applied to it. The authors detail how call-handlers are extensively trained not only handle these calls but also provide effective assistance to the callers. Furthermore, they detail that in a crisis situation, supervisors support the call-handler to offer the best possible assistance. This support can involve reaching out to local authorities, referrals to other agencies and arranging follow-up contact (Cockbain and Tompson, 2024). This is important because it demonstrates that the primary function of these hotlines is to provide help, not merely to collect data. The authors the detail the data pre-processing phases, which involved studying individual entries and setting thresholds for exlucding certain data. After this data cleaning phase their final dataset included: 8,535 contacts in total – predominantly via calls (n=5,213,61.1%), followed by emails (n=2,234, 26.2%) and web forms (n=1,067, 12.5%)" (Cockbain and Tompson, 2024). This discussion of how the data was pre-processed is valuable as it provides me with insights as to how the data was handled before it was delivered to data scientists. It also provides me with various tools to be able to understand handling pre-processing data for my own analyses.

Next, the researchers discuss their Exploratory Data Analysis (EDA), a process involving both multivariate and univariate analyses to gain a better understanding of the dataset. Given the sensitivity of the data, the researchers consulted with the helpline staff for clarifications and to develop theories surrounding their analysis. One analysis provided valuable information about the nature of the callers. Figure 5 illustrates a breakdown of the different types of exploitation reported. It can be seen that the bulk of the callers were victim to Labor Exploitation (52.7%). Another statistic that drew my attention was that 12.9% of the exploitation was unknown, these cases were for individuals that did not want to divulge their situation or were calling simply for information purposes (Cockbain and Tompson, 2024).

Figure 5: *Breakdown of potential exploitation types within the Helpline data by case (n=3,613)*

Exploitation type	*n*	%
Labour exploitation	1,903	52.7
Sexual exploitation	533	14.8
Unknown	465	12.9
Domestic servitude	386	10.6
Criminal exploitation	206	5.7
Various	120	3.3
Total	3,613	100.0

Note. From "The role of helplines in the anti-trafficking space: examining contacts to a major 'modern slavery' hotline" by Ella Cockbain and Lisa Thompson. Crime, Law and Social Change, https://doi.org/10.1007/s10611-024-10151-z

With this EDA, the researchers then delve deeper into each type of exploitation and identify the sector in which the exploitation in occurring. From these analyses, several insights can be drawn. For example, a significant portion of labor exploitation occurs at car washes and restaurants. In the case of criminal exploitation, the majority involves activities related to "other drugs" and cannabis farms (Cockbain and Tompson, 2024). This information is valuable as it can encourage organizations and local authorities to increase surveillance and monitoring of such establishments. Furthermore, it can serve as compelling data for politicians to introduce legislation that protects these marginalized individuals. Figure 6 below shows sector-by-sector breakdown of these exploitation types.

Figure 6: *Cases (n=3,613) reclassified into potential exploitation type, sectors and sub-sectors*

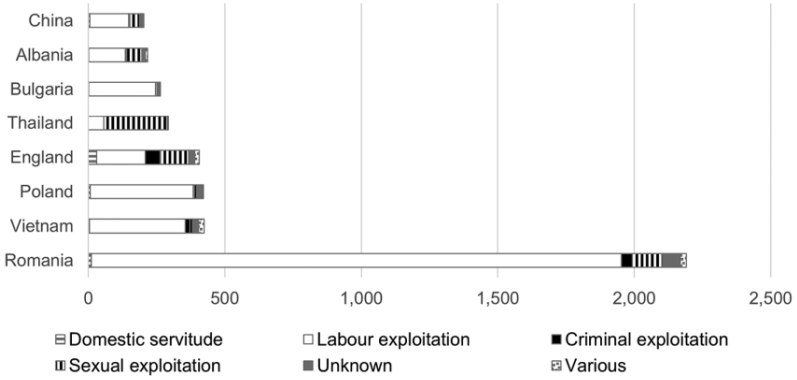
Exploitation category	Exploitation sector category	Exploitation sector sub-category	n
Labour exploitation	Care sector	-	19
	Construction	-	239
	Entertainment	-	11
	Food production	Agriculture / Farm	74
		Factory	4
	Hospitality	Hotel / Motel	28
		Take away / Restaurant	160
		Other	5
	Manufacturing	Factory	48
		Other	15
	Maritime industry / Boat / Shipping	-	11
	Services	Beauty / Spa	221
		Car wash	625
		Shop	14
		Other	42
	Transportation	-	27
	Various	-	136
	Other	-	224
Subtotal			**1,903**
Sexual exploitation	Commercial	Brothel	298
		Hotel / Motel	12
		Private home	4
		Street	17
		Various	97
	Various	-	8
	Other	-	14
Subtotal			**533**
Unknown	Other	-	465
Subtotal			**465**
Domestic servitude	Domestic work / Au pair / Nanny	-	386
Subtotal			**386**
Criminal exploitation	Criminal	Benefit fraud	1
		Cannabis farm	36
		Other drugs	54
		Pickpocketing	1
		Shoplifting	4
		Street begging	82
		Other	20
	Various	-	1
	Other		7
Subtotal			**206**
Various	Various	-	120
Subtotal			**120**
Overall total			**3,613**

Note. From "The role of helplines in the anti-trafficking space: examining contacts to a major 'modern slavery' hotline" by Ella Cockbain and Lisa Thompson. Crime, Law and Social Change, https://doi.org/10.1007/s10611-024-10151-z

To better identify which groups require further assistance the researchers then delved into the gender and nationality of the callers to determine which demographics were most frequent. They concluded that "Disregarding the instances where the sex/gender was unknown (21.4%, n=897), three quarters of potential exploiters were recorded as male (n=2,528), 23% as female, (n=761), 0.09% as transgender or gender non-conforming (n=3)" (Cockbain and Tompson, 2024). When we consider nationalities it can be seen that individuals from Romania are more frequently victim to exploitation more than the other nationalities

listed. However, different categories of exploitation show varied distributions across gender and nationality. A breakdown of the the at-risk nationalities is shown in Figure 7 below.

Figure 7: *Breakdown by exploitation type for the eight most common nationalities (countries) recorded for potential victims at individual level (n=4,419)*



Note. From "The role of helplines in the anti-trafficking space: examining contacts to a major 'modern slavery' hotline" by Ella Cockbain and Lisa Thompson. Crime, Law and Social Change, https://doi.org/10.1007/s10611-024-10151-z

As previously discussed one of the main functions of these helplines is to refer individuals to outside agencies and law enforcement when necessary. The first figure (Figure 8) below illustrates the "onward" actions taken based on each exploitation type. It is important to note that no referrals were made in cases where the exploitation type was unknown due to insufficient information. The authors emphasize that a significant majority of referrals, about 83% , are to law enforcement for further investigation. Figure 9 below shows that calls made on behalf of someone else (i.e. calling for someone) resulted in a higher rate of referral to law enforcement authorities compared to victims themselves (Cockbain and Tompson, 2024). The authors asked the Helpline staff for further guidance as to policy and principles with regards to how they handle referrals and they claim the following:

> They explained their guiding principle is 'to do no further harm', and they seek to establish consent for any referrals from the contact(s) and the person/people potentially being exploited (who may be one and the same). All decisions to refer are reviewed by a second person (in complex cases, a senior manager). The Helpline is not legally required to report to the authorities, they said, and staff are led above all by individual cases' specifics. Broadly speaking, however, if someone self-reports, their case would not be referred to the authorities without their explicit consent unless they were a minor, and/or assessed to be at 'immediate risk of harm', and/ or others involved met these criteria. In such circumstances, the case handler would reportedly tell

them as soon as possible – ideally before sensitive information is disclosed – that they will call the police/other authorities (Cockbain and Tompson, 2024).

Figure 8: *Breakdown of onward action taken by potential exploitation case, by case (n=3,613)*

Exploitation type	Neither referrals nor signposts % (n)	Just signposts % (n)	Just referrals % (n)	Both % (n)	Total % (n)
Labour	*13.5* (257)	*4.0* (76)	*78.8* (1,499)	*3.7* (71)	*100.0* (1,903)
Sexual	*28.9* (154)	*15.4* (82)	*49.7* (265)	*6.0* (32)	*100.0* (533)
Domestic servitude	*34.2* (132)	*14.5* (56)	*41.2* (159)	*10.1* (39)	*100.0* (386)
Criminal	*36.9* (76)	*10.7* (22)	*46.6* (96)	*5.8* (12)	*100.0* (206)
Unknown	*51.8* (241)	*22.6* (105)	*23.9* (111)	*1.7* (8)	*100.0* (465)
Various	*39.2* (47)	*15.0* (18)	*43.3* (52)	*2.5* (3)	*100.0* (120)

Note. From "The role of helplines in the anti-trafficking space: examining contacts to a major 'modern slavery' hotline" by Ella Cockbain and Lisa Thompson. Crime, Law and Social Change, https://doi.org/10.1007/s10611-024-10151-z

Figure 9: *Breakdown of referrals to law enforcement by exploitation case and proximity of caller to victim (n=2,111)*

	Exploitation type											
	Domestic servitude		Labour		Criminal		Sexual		Unknown		Various		Total
Proximity	n	%	n	%	n	%	n	%	n	%	n	%	n
Victim self-report	25	14.2	129	9.1	9.1	6.3	29	11.4	7	5.6	11	20.4	206
Direct contact with potential victim	96	54.5	449	31.6	31.6	34.2	89	34.9	51	41.1	18	33.3	730
Indirect contact with potential victim	15	8.5	114	8.0	8.0	1.3	43	16.9	4	3.2	5	9.3	182
Observation of suspicious activity	37	21.0	687	48.3	48.3	57.0	80	31.4	49	39.5	15	27.8	913
Unknown	3	1.7	44	3.1	3.1	1.3	14	5.5	13	10.5	5	9.3	80
Total	176	100	1,423	100	9.1	100	255	100	124	100	54	100	2,111

Note. From "The role of helplines in the anti-trafficking space: examining contacts to a major 'modern slavery' hotline" by Ella Cockbain and Lisa Thompson. Crime, Law and Social Change, https://doi.org/10.1007/s10611-024-10151-z

Finally, the authors discuss the limitations of this study. They note that some activities may go unnoticed as some activity does not meet the legal definition of the trafficking that is outlined by UK law. Additionally, they highlight systemic biases, as certain individuals may be less willing to contact a helpline. This is directly correlated to some missing information that limited further analysis of more nuanced issues. Additionally, the authors state that the lack of temporal data prevented them from being able to conduct a time-series analysis of the data. Ultimately, the authors hope that this research will encourage further empirical research into the matter specifically with helplines that were involved with Russia's war in Ukraine. The authors believe that having this information will provide an insight into user-experience as well as provide the helpline with a larger picture of how these helplines can better operate. This is precisely the purpose of Objective 2. I am seeking to see how the Telegram and 527 IOM data can be used to identify what individuals are asking for and vulnerable populations so that resources can be appropriately allocated and predictions can be made based off of accurate modeling.

3 About the Datasets

This section highlights the datasets that were used in this study. The 2 datasets that I am looking at are a Telegram dataset and data from the 527 helpline in Ukraine. Telegram is a cloud-based instant messaging service that has gained significant prominence in recent years, particularly in crisis situations. Unlike traditional communication platforms, Telegram offers end-to-end encryption, large group chat capacities, and the ability to broadcast messages to an unlimited number of subscribers through channels. These features make it a vital tool for disseminating information quickly and securely during crises. The 527 hotline is a dedicated communication channel established to support Ukrainian citizens, particularly those facing challenges related to migration, displacement, and protection of their rights. Operated by organizations such as the International Organization for Migration (IOM), the 527 hotline provides a range of services, including legal advice, information on asylum procedures, and assistance with locating missing family members. This hotline is especially crucial during times of conflict or humanitarian crisis, as it offers a direct line of support for individuals in distress. By facilitating access to vital information and resources, the 527 hotline plays a key role in safeguarding the rights and well-being of Ukrainians both within the country and abroad. Its establishment and operation highlight the importance of responsive and accessible communication infrastructures in addressing the needs of vulnerable populations during crises.

The Telegram Dataset has 324704 data entries spanning from 09/2022 to 07/2023. On the other hand, the 527 Data has over 88,321 entries spanning from 09/2021 to 09/2023. One important fact that I would like to highlight is that although there is more data in the Telegram dataset the 527 data is more well organized and processed than the Telegram data. First, as part of National Science Foundation Award 2330311 D-ISN/RAPID: Data Collection for Human Trafficking Recruitment and Responses in Forced Migration, Telegram data was scraped from groups that were looking to facilitate information exchange, provide contacts and help for individuals that were trying to relocate within Ukraine as well as individuals that were trying to migrate out of the country. Naturally, this data was in several languages predominately in Ukrainian, but also in Romanian, Russian, and Polish. Thus, one of the first tasks was translating the accurately. WPI graduate students (Amir Jamali and Solomiya Sorokotyaha) translated the scrapped data into English using a combination of translation scripts and manual human verification to generate an accurate translated file. The next step involved parsing these messages to ensure no empty entries in the messages column, minimizing issues with missing entries in subsequent analyses. Another challenge, was handling emojis, which cannot be parsed in a Natural Language Processing (NLP) analyses. These emojis were converted to a text format. For example, (!!) was changed to double_exclamation_point in the messages. A significant concern was the

presence of suspicious messages, within the group, either offering help or were asking for egregious favors. To automate identification of suspicious messages, we collaborated with Dr. Laura Dean from Millikin University. Dr. Dean is the foremost expert on human trafficking in Eastern Europe. With Dr. Dean, we identified emojis commonly associated with trafficking and used n-grams of known suspicious messages to parse out these messages as effectively as possible. Despite htese efforts, some suspicious messages may have slipped through the parsing potentially impacting the analysis.

4 Methods

4.1 Telegram Dataset - Examine social networks on social media

4.1.1 Exploratory Data Analysis (EDA)

The first important step to any data analysis process is to conduct a Exploratory Data Analysis (EDA) to gain a deeper understanding of the dataset. This is particularly valuable there is minimal prior knowledge of the dataset. The following section outlines the EDA process for the Telegram dataset, and explains the rationale behind the chosen methods.

Initially, a histogram for the average number of words was considered. According to a recent study, histograms especially for language-based datasets, can provide valuable insights into the nature of the data (Reif et al., 2024). Such insights can include identifying the frequency of messages, detecting outliers, and a clearer idea of how to proceed with further analysis. As the next step in the EDA, a histogram of stop-words and non stop-words was considered. Before we delve into this, it is important to address *What exactly are stop-words?* A recent study conducted on the influence of stop-words identified that words such as "a", "the", "is", "to and etc. are considered to be noise in the data that can reduce the quality of subsequent analyses (Munková et al., 2014). Furthermore, the authors of the study assert that the removal of these stop-words does not affect the quality or quantity of the textual data being pre-processed (Munková et al., 2014). Moving forward, we analyzed histograms of these stop-words to identify unnecessary words and determine the reduction in word count within the entire dataset. Additionally,we considered histograms of the "top n-grams" . N-grams provide a straightforward way to convert text data into numerical features that machine learning algorithms can process. For instance, unigram (1-gram), bigram (2-gram), and trigram (3-gram) models can represent text in a structured format. N-grams are also valuable for feature extraction by identifying frequent phrases or collocations that are meaningful. This is useful in various tasks like text classification, sentiment analysis, and information retrieval, the first of which is the main goal of a subsequent Machine Learning Analysis (Jurafsky and Martin, 2009).

To summarize, the EDA procedure for the Telegram dataset entailed using histograms to examine various aspects of the data to gain a preliminary understanding. By analyzing the average word count, the proportion of stop-words to non-stop-words, and the most common n-grams, we can better understand the structure of the dataset and pinpoint possible topics for additional research . This initial stage is crucial because it creates the framework for later machine learning assignments and guarantees that the data is

clear, pertinent, and well represented for sophisticated analytical methods.

4.1.2 Machine Learning Analysis with Latent Dirichlet Allocation (LDA)

The goal for our Machine Learning analysis is to group the messages in the dataset to identify clusters of messages related to what the users in the group are requesting or offering. before processing, it is important to develop a brief understanding of what Topic Modeling with LDA (Kelechava, 2020). In Natural Language Processing (NLP), topic modeling seeks to uncover hidden semantic structure within written materials. These probabilistic models assist in sorting through enormous volumes of unprocessed text, automatically grouping similar documents together. , Latent Dirichlet Allocation (LDA), a technique initially proposed in 2000 and later independently developed by by Andrew Ng in 2003, is a key method used in topic modeling. According to LDA, each document in a corpus is a mixture of a predefined number of topics. Each topic has an equal chance of producing a variety of words, where the words are all the words that are noticed in the corpus. Then, depending on the chance of word co-occurrence, these "hidden" subjects are revealed. This is a formal case of a Bayesian inference problem (Blei et al., 2001).

The LDA code used to create and visualize topic models from text data is provided in the Appendix A. The code that I have written adheres to a standardized procedure for topic modeling. The **get_lda_objects** function first prepares the data by applying a series of pre-processing steps and downloading English stop-words. Each document in the text corpus undergoes tokenization, stop-word removal, lemmatization, and filtering to preserve words longer than two characters. The Gensim library's Dictionary class is then used to transform the processed text into a bag-of-words (BoW) representation, producing a dictionary and matching BoW corpus. Using this BoW corpus as a training set, the LDA model is trained with eight topics (from the predefined 9), ten passes (standard), and four worker threads to take advantage of multicore processing. The function returns the dictionary, BoW corpus, and trained LDA model . These outputs are used by the **plot_lda_vis** function and utilizes the pyLDAvis library to create an interactive visualization, which is then rendered in a Jupyter notebook. This visualization helps in exploring and interpreting the topics generated by the LDA model. Finally, the code applies these functions to the dataset's **message_english_cleared** column to produce the topic visualization.

4.2 527 IOM Dataset: Impact of Full-Scale Invasion on Migrants and Internally Displaced Persons

4.2.1 Exploratory Data Analysis (EDA)

For the 527 IOM Data (https://ukraine.iom.int/), a dataset with over 88,000 entries, the EDA was an important process as it provided valuable insights that can help the helpline improve their efforts to assist vulnerable Ukrainians. It is important to note that some assumptions were made as the helpline due to the lack of an adequate data dictionary . With the help of Dr. Laura Dean, we were reached a consensus as to how we should move forward with this data. In this EDA, we employed a variety of graphics, such as histograms, to gain crucial insights into the dataset, which can can be used for future analyses (Reif et al., 2024). Histograms, in particular, are valuable tools for large datasets as they reveal important patterns and trends that inform subsequent analytical steps. Next, for certain categorical variables, such as for Type of Consults and Immigration Category, pie charts are used. Pie charts are an excellent tool for representing differences among variables especially in larger datasets, , as they provide a clear visual comparison of the proportions. (Juggins and Telford, 2012). First and foremost, the number of calls over time is plotted on a histogram to observe how the call volume evolved before and after the invasion. Similarly, we examine histograms for the following variables to analyze their changes, over time particularly before and after the invasion. These are the columns that are considered: **Attitude towards departure, Duration of call, Crossed the border without a visa, Refugee Status Over Time, and Employment and Education**.

Within these columns, we generated specific graphics for certain age-groups and genders to provide the helpline with valuable information regarding high-risk groups or those requiring more attention. For example, we consider border crossings without a visa for males ages 18-65 as this is the draft age, and many men were leaving the country to avoid being drafted. We consider Refugee Status for both men and women to identify the most frequent and least frequent callers. Additionally, we considered Education and Employment to understand the occupation and education level of those calling. It is important to note that the nature of this study is to not criticize or judge the education and employment of individuals. instead, it aims to provide these groups with the adequate support they need to overcome challenges they are encountering.

Lastly, the pie charts were created based on the aforementioned variables to provide the helpline with valuable information on specific groups of people and how these categorical variables shift with among them. For example, we analyzed how different age groups found out about the helpline, the main subject of the

call in the entire dataset, and specifically for Internally Displaced Persons (IDPs) and Potential Migrants. Collectively, these graphics provide us with valuable insights into the data, helping to identify trends and areas where support is most needed.

In conclusion, the EDA of the 527 IOM Data, encompassing over 88,000 entries, has proven to be an invaluable process in understanding and improving the helpline's efforts to assist vulnerable Ukrainians. Despite initial challenges due to the lack of clear definitions from the helpline, collaborative efforts with Dr. Laura Dean allowed us to make informed decisions on data interpretation. By employing a variety of graphical methods, such as histograms and pie charts, we extracted crucial insights into various aspects of the dataset, including the number of calls over time (Figure 16) and refugee status (Figure 19). These visualizations not only highlighted trends and patterns within the data but also identified high-risk groups and areas needing more support. Ultimately, this analysis offers a comprehensive overview of the dataset, enabling the helpline to tailor their interventions more effectively and support vulnerable Ukrainians during these challenging times.

4.2.2 Regression Model

The dataset used in this study has a number of characteristics pertaining to the defense of Ukrainian nationals' rights overseas. **"Protection of the rights of Ukrainians abroad"** was chosen as target variable for the the regression model. Preprocessing was done to accommodate both numerical and categorical features. Initially, all categorical columns were converted to string format to ensure they would work with the preprocessing stages. The **select_dtypes** function was used to identify the categorical columns, which were subsequently translated to strings. A pipeline for categorical data was built to manage missing values and encode categorical features. This pipeline consisted of two steps: one-hot encoding to transform categorical data into a format suitable for the regression model, and imputation of missing values using a constant method **('missing')**. Another pipeline was created for numerical features. This pipeline comprised Principal Component Analysis (PCA) for dimensionality reduction, **StandardScaler** for feature scaling, and the mean method for imputation of missing values. The inclusion of PCA allowed for automatic dimensionality reduction based on the features of the dataset, without specifying the number of components. The **ColumnTransformer** was used to combine these preprocessing procedures, applying the appropriate transformations to both the numerical and categorical columns of the dataset.

The preprocessed dataset was then split into training and testing sets using an 80-20 split to ensure the model was trained on a substantial portion of the data while keeping a significant portion for evaluation.

To ensure reproducibility of results, the split was executed using the **train_test_split** function with a fixed random state of 42.

The Random Forest Regressor was chosen for the regression model due to its resilience and versatility. To guarantee consistency in the outcomes, the model was defined using 100 estimators and a fixed random state of 42. The model was defined using 100 estimators to ensure a robust and stable prediction by averaging over multiple decision trees, which reduces variance, and a fixed random state of 42 to ensure reproducibility of the results by controlling the randomness in data splitting and model training. The regression model and the preprocessing steps were combined into a single pipeline. The model was then trained by fitting this pipeline to the training set **(X_train and y_train)**. Predictions were produced using the test data **(X_test)** following training. To ensure accurate assessment, any missing values in the test data and predictions were addressed by substituting NaN for "missing" entries, which were subsequently imputed using the mean technique. This imputation was necessary to avoid errors in the calculation of evaluation metrics . The model's performance was evaluated using Two measures: the Mean Squared Error (MSE) and the R-squared (R2) score. The R2 score calculates the percentage of the dependent variable's variation that can be predicted based on the independent variables, whereas the MSE shows the average squared difference between the anticipated and actual values.

A residuals plot was created to evaluate the regression model's performance in more detail. This plot helps identify trends or anomalies in the model's predictions by displaying the difference between the observed and projected values. A scatter plot of the residuals against the expected values was made, with the ideal scenario – where the predictions exactly match the observed values – represented by a horizontal line at zero. Figure 27 is essential for identifying possible model problems, such as heteroscedasticity, non-linearity, or outliers. In addition to assessing the model's functionality, an analysis of the most prominent aspects was conducted to comprehend their significance to the model. The Random Forest Regressor, which provides a score indicating each feature's usefulness in predicting the target variable, was used to determine each feature's significance. The key characteristics were noted along with an explanation of their importance. This analysis is important because it identifies the main variables that affect how well Ukrainians' rights are protected overseas and provides guidance on which of these variables should take precedence in future studies or policy decisions.

5 Results

Our analysis provides comprehensive insights into two important datasets: the 527 IOM Dataset and the Telegram dataset. Despite their differences, these datasets complement each other, enhancing our understanding of the Ukrainian context from multiple perspectives. The Telegram dataset, with its collection of user-generated communications, offers a nuanced view of public opinion and discourse about the ongoing crises. By applying topic modeling techniques and exploratory data analysis (EDA), we can identify recurring themes and patterns in communication, which provide insight into the attitudes and worries of the general public. On the other hand, the 527 IOM Data, with over 88,000 entries, captures structured data on helpline calls to the helpline, reflecting the needs and real-world challenges of those affected by the conflict. We carefully examine various aspects of this dataset, including the type and frequency of calls, demographic information, and specific issues encountered by certain groups. Additionally, we use a regression model to see if Ukrainian rights are being protected. By adopting a dual approach – analyzing both the qualitative and quantitative insights from the Telegram dataset and the quantitative data from the IOM dataset - we can enhance our understanding from multiple perspectives, helping to improve support systems for those Ukrainians affected by the conflict.

In the following sections, we detail specific results of our analysis of these datasets. First, we investigate the Telegram data, highlighting the major themes identified by topic modeling as well as the insights gained from sentiment and message frequency analysis. Next, we examine the IOM Data, focusing on call pattern trends, demographic information, and the specific needs of high-risk populations. When combined, these findings provide a thorough picture of the state of affairs and practical insights for helping individuals impacted by the conflict.

5.1 Results for the Telegram Dataset

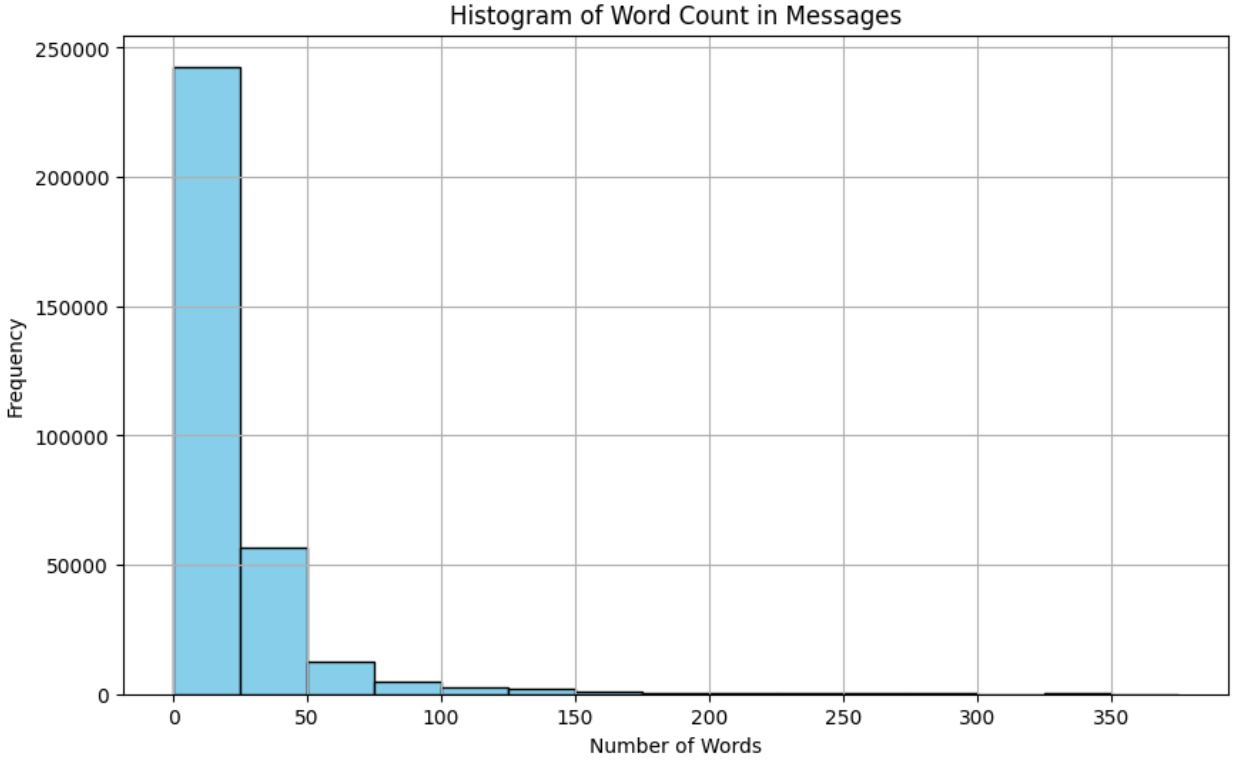
5.1.1 Exploratory Data Analysis (EDA) Results

Now we delve into the results for the EDA. First, we had considered a histogram of the all the messages and examined word frequencies for messages across the entire dataset. As seen in the figure below (FIGURE 10), the majority of the messages fell into the 0-50 word group, totalling approximately 299,261 messages in total.

What could this mean? It appears that most messages are in this group because upon reviewing the actual content, the majority of the messages are concise, often asking for quick information pertaining to

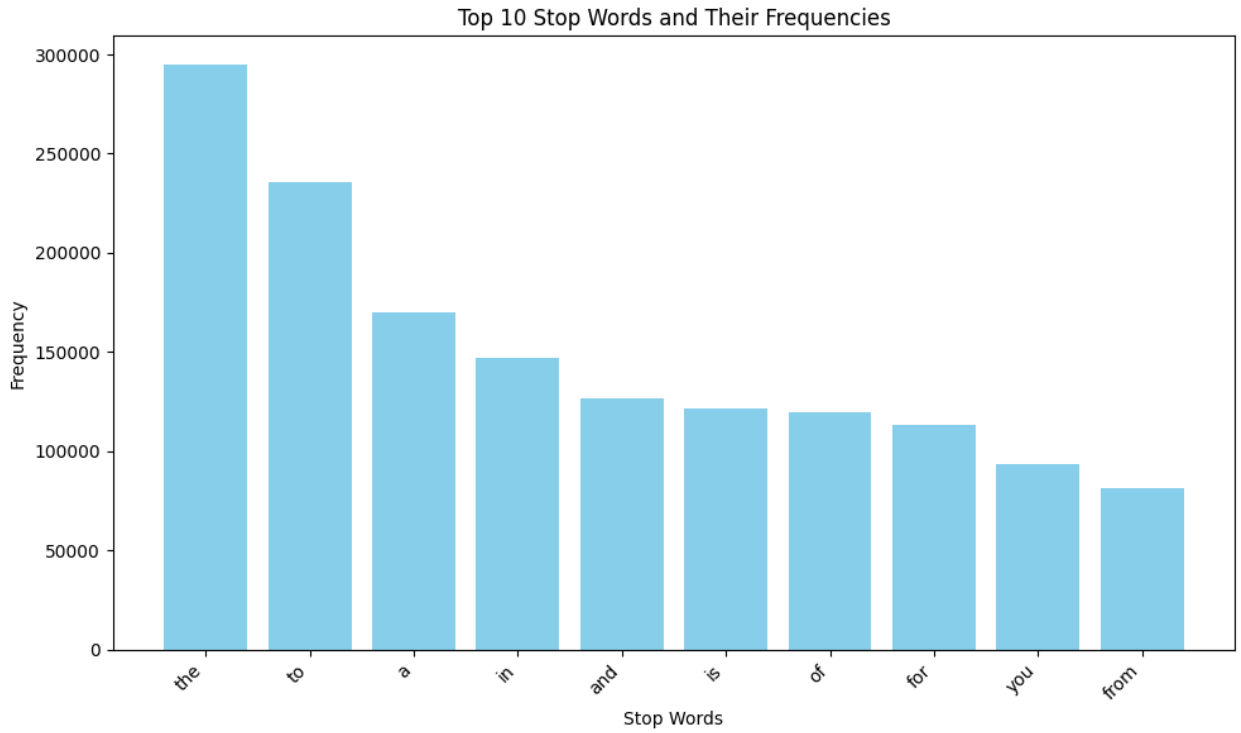
their specific goal. However, it is also important to consider the longer messages. The data shows that there are 342 messages exceeding 300 words. Upon reviewing the original messages, these entries tend typically request detailed information pertaining to transportation and asking for legal guidelines for crossing the border.

Figure 10: *Histogram of Word Count in the Messages*



Next, examined a histogram of stop-words. Analyzing stop-words is valuable as it provides insights into how many of the words in the messages are in this category. In language-based datasets, stop-word analysis is crucial, especially in the field of natural language processing (NLP). Stop-words are everyday words that are used frequently in a language but don't convey much important meaning. Stop-words contain words like "the," "is," "in," and "at." By detecting and eliminating stop-words from a dataset, we can focus on the most important words that enhance the comprehension and analysis of textual material. This process benefits various NLP activities, such as text categorization, sentiment analysis, and information retrieval. Additionally, removing stop-words helps in reducing the dimensionality of the data, thereby improving the functionality of machine learning algorithms. It also reduces noise, , ensuring that the analysis is not skewed by the overwhelming presence of these common words. Based on the figure below (FIGURE 11), the most common stop-words in this dataset are "the", "to" and "a". Given this information we next look at the most-common non-stop-words that are in the dataset.

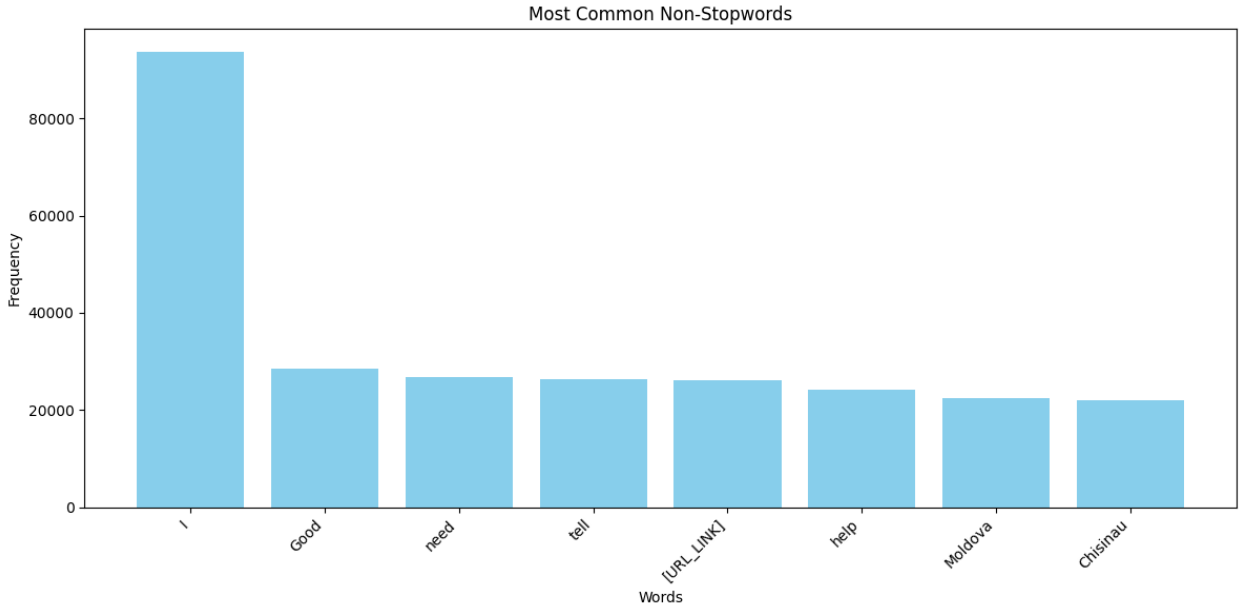
Figure 11: *Top 10 stop-words in the dataset*



After analyzing the stop-words, we then examined the non-stop-words present in the dataset. When working with language-based datasets, it is crucial to look at non-stop-words, also known as content words, as these words convey the main idea and context of the text. Unlike stop-words, which are frequently used and have minimal semantic significance, non-stop-words include nouns, verbs, adjectives, and adverbs that are essential for comprehending the topic and tone of the text. By concentrating on non-stop-words, researchers can obtain a greater understanding of the themes, issues, and nuances found in the data.

Non-stop-words facilitate the identification of important ideas and connections within the text, making analysis more precise and efficient. Furthermore, by prioritizing non-stop-words, analysts can enhance the efficacy of machine learning models, as these terms offer the discriminative attributes required for accurate prediction and classification. Given this information. Based on the figure below (FIGURE 12), the most popular non-stop-words are "I", "Good" and "need". Based on these results, the prevalence of "I" can be attributed to individuals asking for themselves or their families, while "need" can be reflects the fact that the bulk of the entries in this dataset are asking for help. Additionally, the presence of [URL_LINK] is noteworthy, as it indicates that people are directing individuals to external sites either for help or for resources that can better help the individual.

Figure 12: *Top 10 non-stop-words in the dataset*



Following the analysis of the stop-words, I examined the n-grams in the dataset, specifically the tri-grams and quad-grams. Analyzing n-grams is crucial in language-based datasets, as they capture the sequential patterns and contextual relationships between words. An n-gram is a contiguous sequence of n items from a given sample of text or speech, where n can be any integer. By examining n-grams, researchers can determine and measure the frequency of word combinations, revealing language structures and popular phrases visible that are not apparent when examining individual words. N-grams help differentiate various word contexts, thereby enhancing the accuracy of language models and predictions. Furthermore, they caputure idiomatic phrases, collocations, and other multi-word constructions that are critical for a fuller comprehension of the text can be captured by n-grams.

Consequently, n-gram analysis is an essential component of natural language processing, offering deep insights into the syntactic and semantic characteristics of language-based datasets and improving textual data analysis quality overall. In the figure below (FIGURE 13), we can see the most popular tri-grams and quad-grams in the data. It appears that most of the n-grams are requests for help, with some entries offering assistance, as indicated by phrases such as "it possible to" and "it is possible to" in the n-gram graphics. Additionally, it is also important to note the presence of emojis. A review of the original messages suggests that people asking for help use emojis to draw attention, while those offering include emojis to provide valuable help and information.

Figure 13: *Top tri-grams*

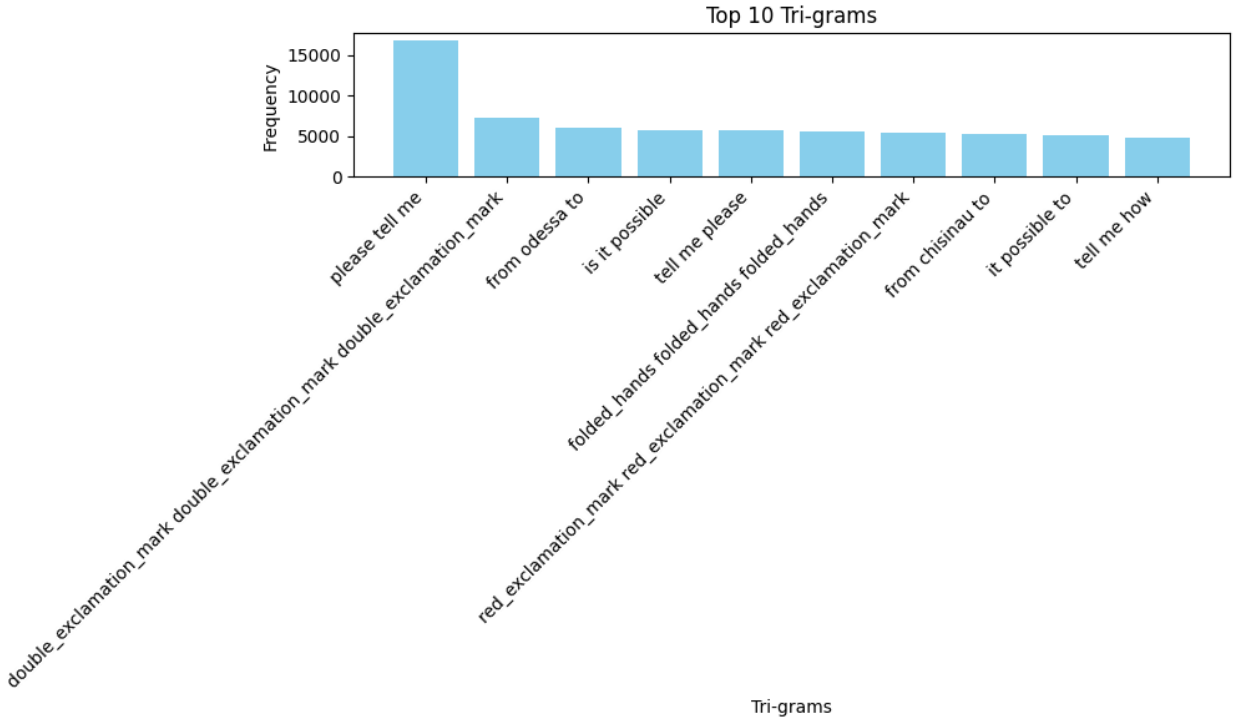
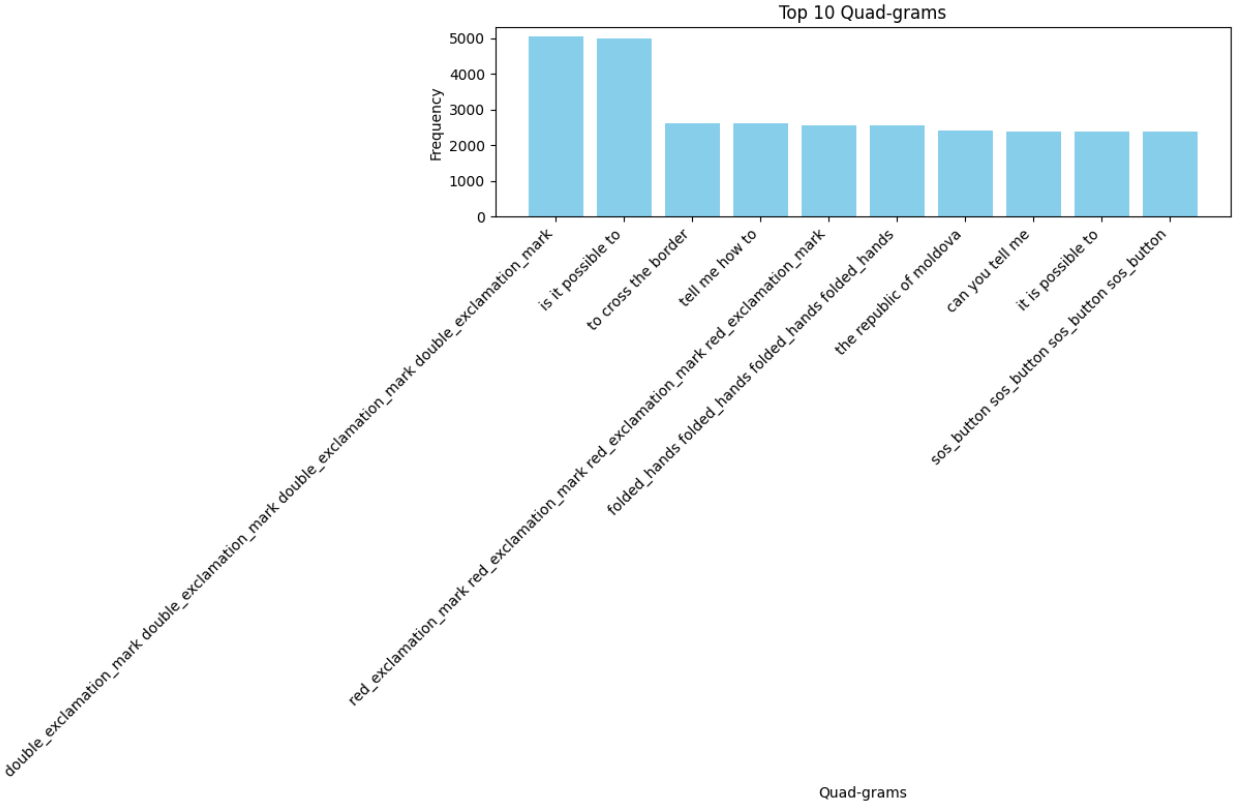


Figure 14: *Top quad-grams*



5.1.2 Latent Dirichlet Allocation (LDA) Results

Now, I discuss the results of the LDA algorithm. In the field of topic modeling for language-based datasets, Latent Dirichlet Allocation (LDA) is a crucial method for revealing underlying thematic structures in textual data. LDA operates on the assumption that topics are distributions over words and documents are mixes of topics, making it a probabilistic generative model. Using this model, researchers can identify latent topics that frequently appear within a group of documents. This makes it possible to explore and organize vast amounts of text according to common themes and ideas. LDA is particularly beneficial for exploratory data analysis and content summarization tasks, as it enables analysts to automatically find topics without any prior understanding of the content of the dataset. Additionally, by grouping related documents and identifying key information, LDA simplifies the organization and retrieval of textual data, highlighting themes that speak to specific data subsets. This capability is vital for applications such as recommendation engines, social media content analysis, and academic literature retrieval. Ultimately, LDA improves language-based datasets' interpretability and usefulness by offering an organized framework for analyzing and deriving significant conclusions from textual data.

In the figure below (Figure 15), we see that there are 8 topics. This number 8 was determined based on prior analysis of the groups based on manual reading of the messages that were present in the dataset. These 8 groups were considered when the algorithm was run. Based on the model below here are the groups, (1) No Need Identified, (2) Looking for legal information about border crossings, (3) Request for Transportation, (4) Job Opportunities, (5) Looking for Housing, (6) General Messages, (7) Financial Assistance, and (8) Transportation to another country.

It is also important to address the axis labels PC1 and PC2. PC1 and PC2 refer to the first and second principal components that are derived from the topic-term distribution-matrix. These components are mainly used to represent the topics in a lower-dimensional space for visualization purposes. Specifically, PC1 represents the primary direction of variation in the distribution of terms across topics, capturing the most salient patterns or themes that differentiate topics from each other. PC2 on the other hand, represents the secondary direction of variation orthogonal to PC1, capturing additional patterns or themes that contribute to the differentiation of topics beyond what is captured by PC1. Together, PC1 and PC2 provide a two-dimensional representation of the topics and terms, facilitating easy visualization and interpretation of the topic structure and relationships.

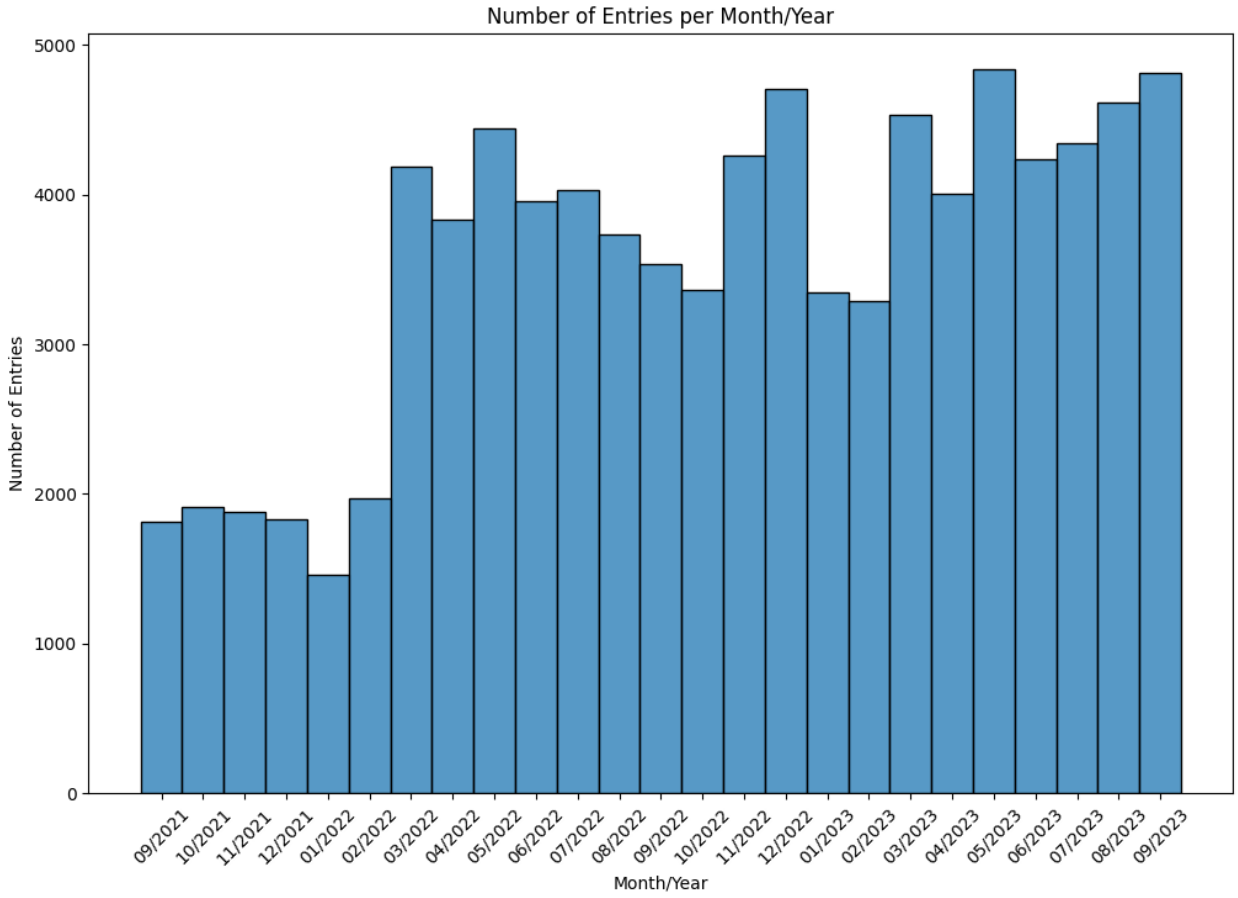
Figure 15: *Results of the LDA, each bubble represents a different topic*

5.2 Results for the 527 Dataset

5.2.1 Exploratory Data Analysis (EDA) Results

Along with the Telegram dataset, I worked with the 527 dataset, a helpline offering help to Ukrainian individuals. This dataset has over 80,000 entries. First, we examined a histogram of the entries over time. As shown in the figure below (FIGURE 16), there is a significant increase in the number of calls received by the helpline after February 2022, jumping from around 2,000 calls to more than 4,000 per month. This can be attributed to the Russian full-scale invasion of Ukraine which began in February 2022. The histogram also indicates that a large number of messages are post-invasion as many individuals in Ukraine were either trying to leave the country or escape conflict areas..

Figure 16: *Histogram of the number of calls from 09/2021 to 09/2023*



Another graphic considered was a pie-chart of the "Type of Consults" that the helpline handled. As seen in the figure below (Figure 17), the majority of consults were informational, or a combination of legal and informational, with a smaller number of calls seeking legal consults. When paired with the histogram of the length of calls which is seen in Figure 18 below, it becomes clear why a large portion of the calls are for informational purposes. Most calls are under 5 minutes, which is understandable that in a state of war most people are on the move seeking quick information pertaining to their movements. This information support the data from both graphics. Furthermore, Figure 18 shows a sharp increase in calls of all lengths after the invasion (02/2022). One question that does arise is *Why are there calls of different lengths?* According to Dr. Dean the longer calls can be attributed to cases that are more of concern and require more help than the average less than 5 minute calls which are asking for quick information.

Figure 17: *Pie-chart of the Type of Consults from 09/2021 to 09/2023*

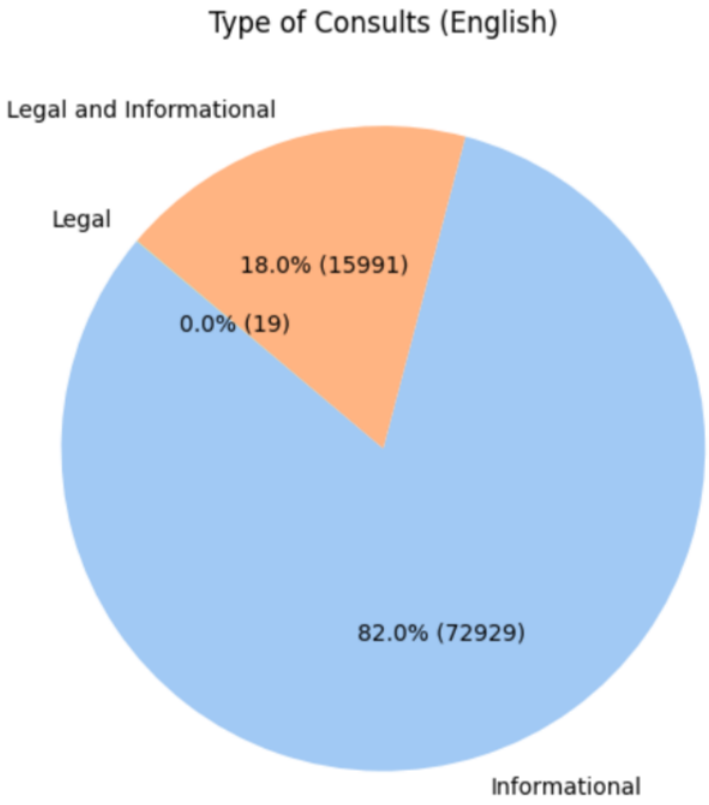
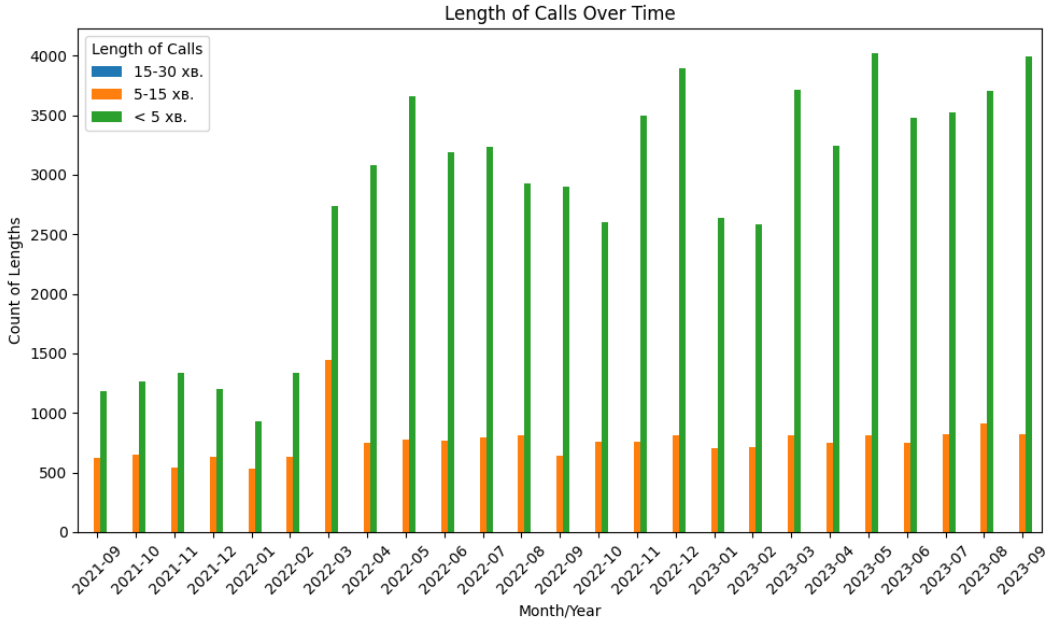


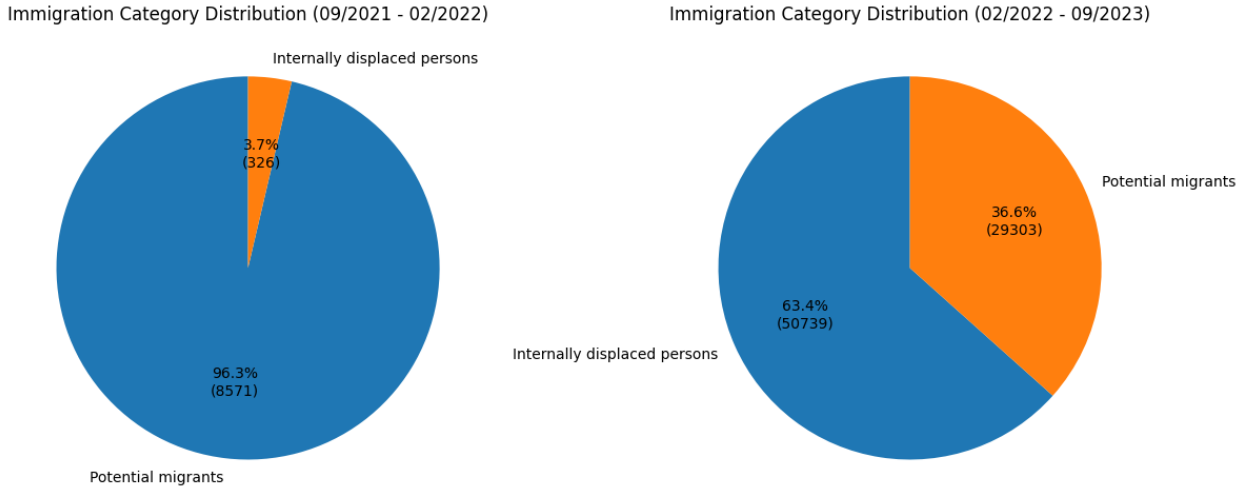
Figure 18: *Histogram of the Call Duration from 09/2021 to 09/2023*



Subsequently, the category of the callers before and after the invasion is considered. By "category of the callers", we refer to their immigration status –either Potential Migrant or an Internally Displaced Person (IDP) (see Figure 19). Analyzing a graphic of Migrant Status, specifically focusing on IDPs and Potential Migrants, before and after the Russian full-scale invasion, is crucial for understanding the impact of this significant geopolitical event on population movements. By comparing migrant statuses prior to and following the invasion, we can identify trends and changes in the proportion of the population that has been internally displaced versus those considering or undergoing migration to other areas or countries. This analysis not only highlights immediate humanitarian effects of the conflict, but also aids in the development and implementation of efficient relief and support initiatives.

Furthermore, providing information about the scope and character of displacement, helps international organizations and policymakers undertake more timely and focused interventions. By incorporating empirical data into the larger research narrative, this comparative graphic effectively illustrates the significant socio-political ramifications of the invasion. In the figure below (see Figure 19), it is evident that before the invasion, the helpline received more calls from Potential Migrants. However, after the invasion, the number of calls from IDPs increased significantly. This can be attributed to people were trying to leave the areas of the country where active conflict was occurring. It is important to note that prior to the full-scale invasion, Ukraine did have IDP from the two eastern oblasts Luhansk and Donetsk which Russia invaded in 2014.

Figure 19: *Pie Chart of Immigration Category from before and after the invasion (09/2022)*



to comprehend the reach and accessibility of support services across various age groups, it is important to consider an Age Distribution graphic and how older adults learned about the helpline. Older people may have different demands and face unique challenges in accessing information compared to younger populations.

By examining how individuals over a specific age, such as forty, found the helpline, support workers can identify the most effective methods communication. This data is essential for customizing outreach strategies to ensure older adults are well-informed of the assistance resources that are accessible to them. A graphic representation of age distribution, segmented into specific age periods, provides a clear and comprehensive depiction of trends (see Figure 20). This enables a detailed examination of whether certain age groups, particularly older adults, are underrepresented in helpline usages, potentially indicating accessibility or communication issues. By comparing this data with general age demographics, we can evaluate the success of ongoing outreach initiatives and identify gaps. Such insights are instrumental for improving the design and implementation of information dissemination strategies. They ensure that support services are inclusive and accessible to all age groups, particularly older individuals who might rely more on traditional media or word-of-mouth. In an academic context, this graphic not only enhances the understanding of demographic-specific outreach but also contributes to the broader discourse on public health communication and support system accessibility.

Figure 20: *Age Distribution of Callers*

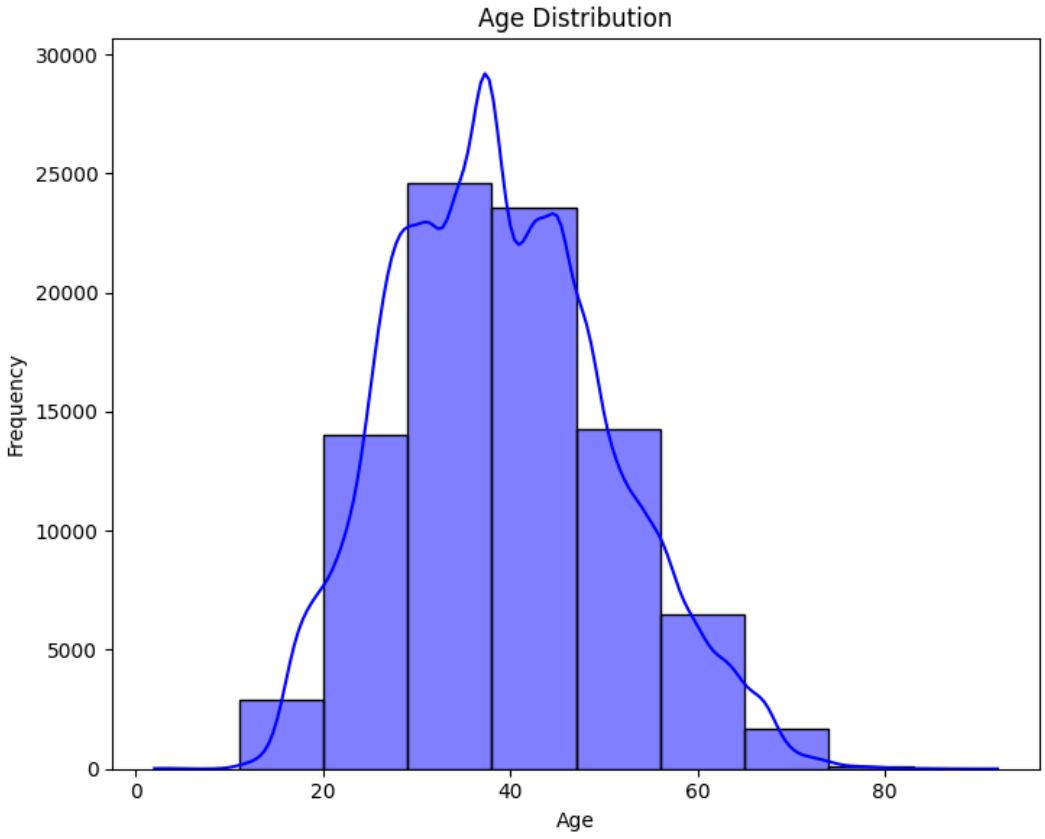
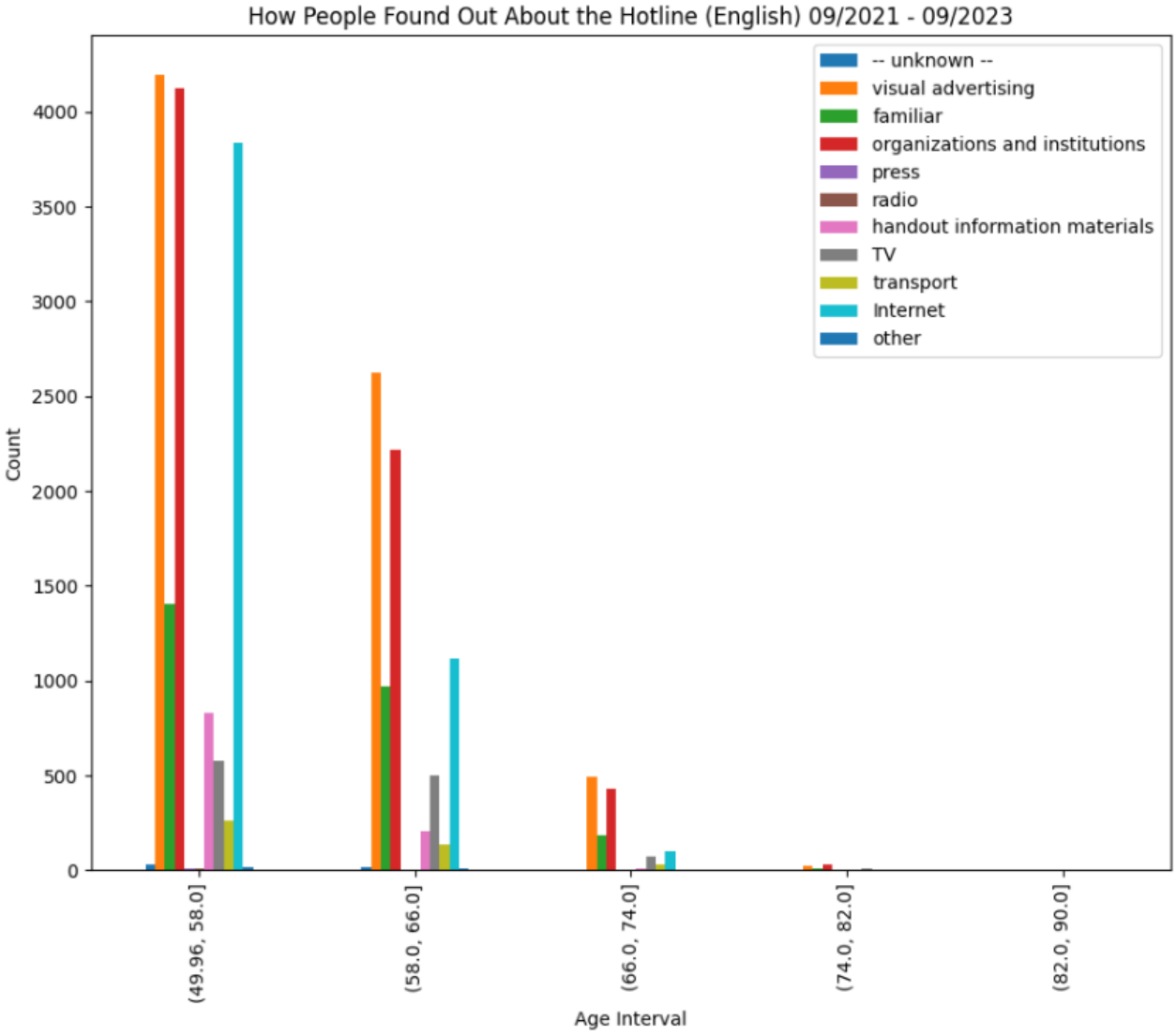


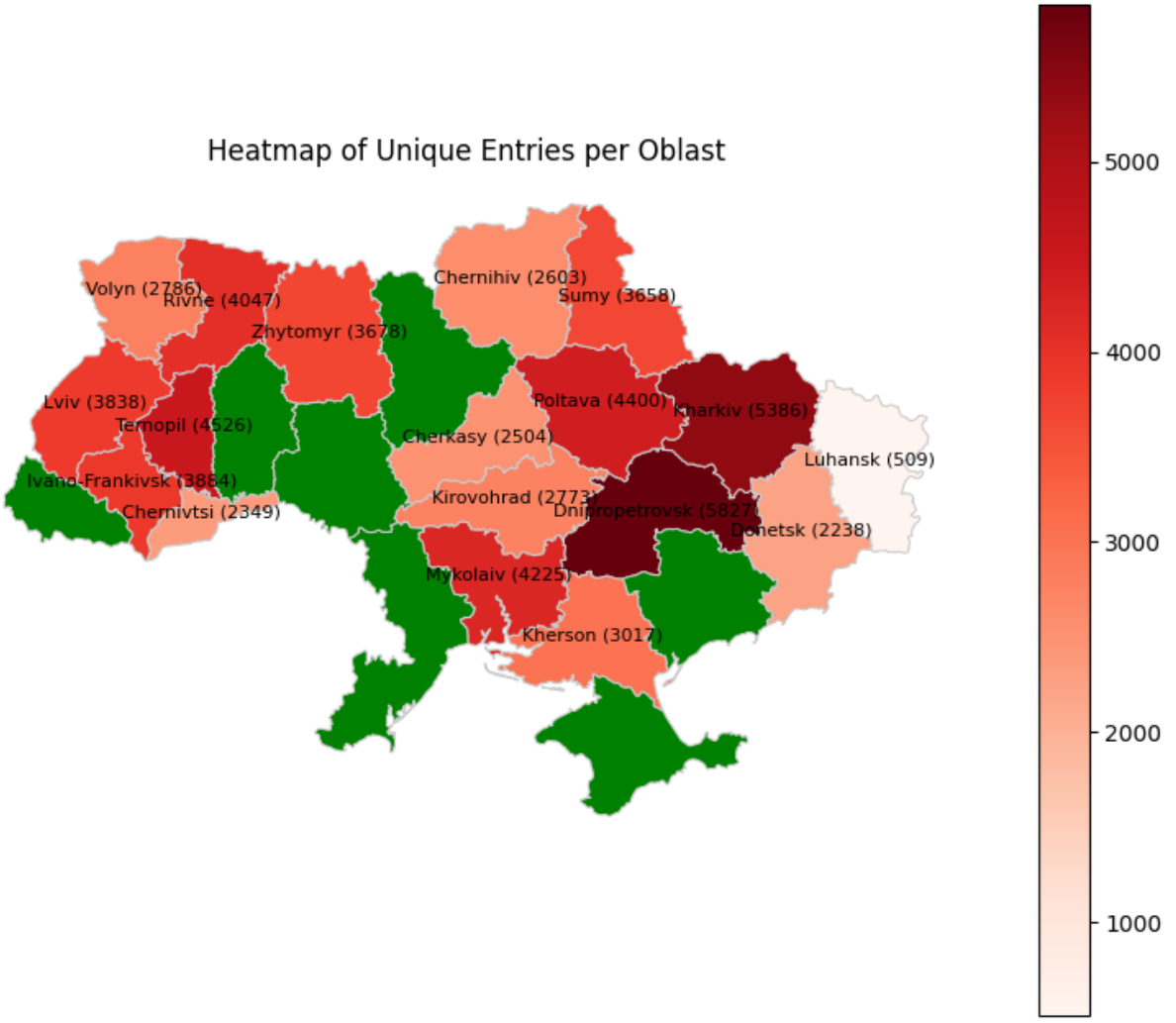
Figure 21: *How Older people found out about the hotline*



To gain a spatial understanding of the helpline's reach and recognize regional patterns in the demand for support services, it is essential to consider a map graphic that shows the regions from which the calls originate. This visualization enables hotline staff and policymakers to identify places with high call volumes, which may indicate severe distress and greater need for assistance. By mapping the call origins, we can evaluate the geographic distribution of the helpline's users, highlighting regions where the service is most frequently used and possibly underutilized. A graphic like the one below (Figure 22) is very helpful in highlighting geographical disparities in the helpline's accessibility and awareness. It can reveal whether particular areas—particularly isolated or rural ones—are aware of the helpline and have access to its services. This information is crucial for developing targeted outreach and intervention strategies to ensure support is

distributed equitably and is available to all regions, particularly those that are marginalized or under-served. Furthermore, mapping call data can shed light on the relationship between local socioeconomic circumstances and the need for hotline services. Higher call volumes, for instance, may indicate more serious socioeconomic problems in the area, necessitating the provision of more resources and assistance. This geographic analysis offers a comprehensive understanding of how regional characteristics influence the consumption of support services, which enhances our overall understanding of the impact and effectiveness of the hotline.

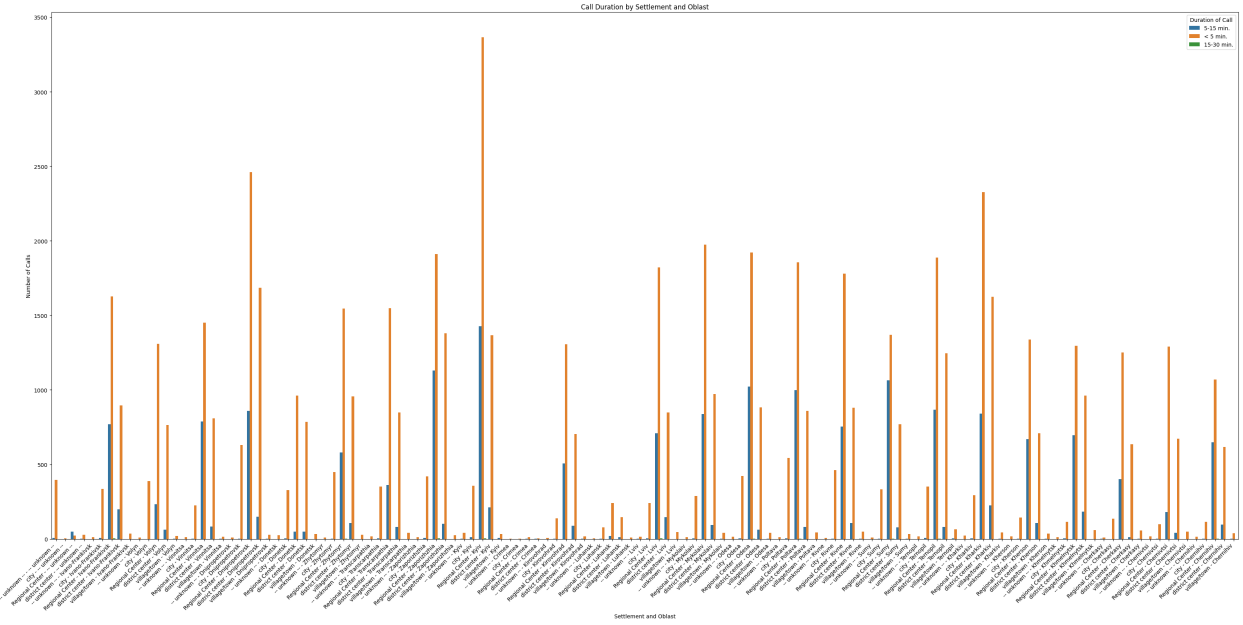
Figure 22: *Heatmap of how many calls are received from each Oblast (09/2021 - 09/2023)*



Understanding the geographical and contextual aspects that affect the length of contacts with the helpline requires a graphic showing the relationship between Settlement, Province, and Length of Calls. This analysis can offer insights into the types of issues people in various areas face and how those problems affect the duration of their calls. (see Figure 23 for more clear viewing see Appendix D)

By mapping the duration of calls across different provinces(oblasts in Ukraine) and settlements, support staff can identify patterns indicating areas that need lengthier consultations due to more complex or severe problems. For example, an oblast that has longer average call durations may be facing more serious socioeconomic problems or higher levels of distress upon its citizens. Such as, Kyiv and Kharkiv which have longer average call durations, and connecting them with events in the real war these were areas where there was heavy bombings and mass migrations. It is also important to note that these are also some more populated regions in the country which could be another reason for the frequency of calls. Conversely, regions with fewer calls might indicate either less complex issues or that the helpline's resources and outreach are more effective in those regions. this can also indicate that the marketing for the 527 hotline is not as effective in this region, meaning that if an area is under occupation, maybe the helpline information isn't reaching target audiences.

Figure 23: *Call Duration by Settlement and Oblast (09/2021 - 09/2023)*

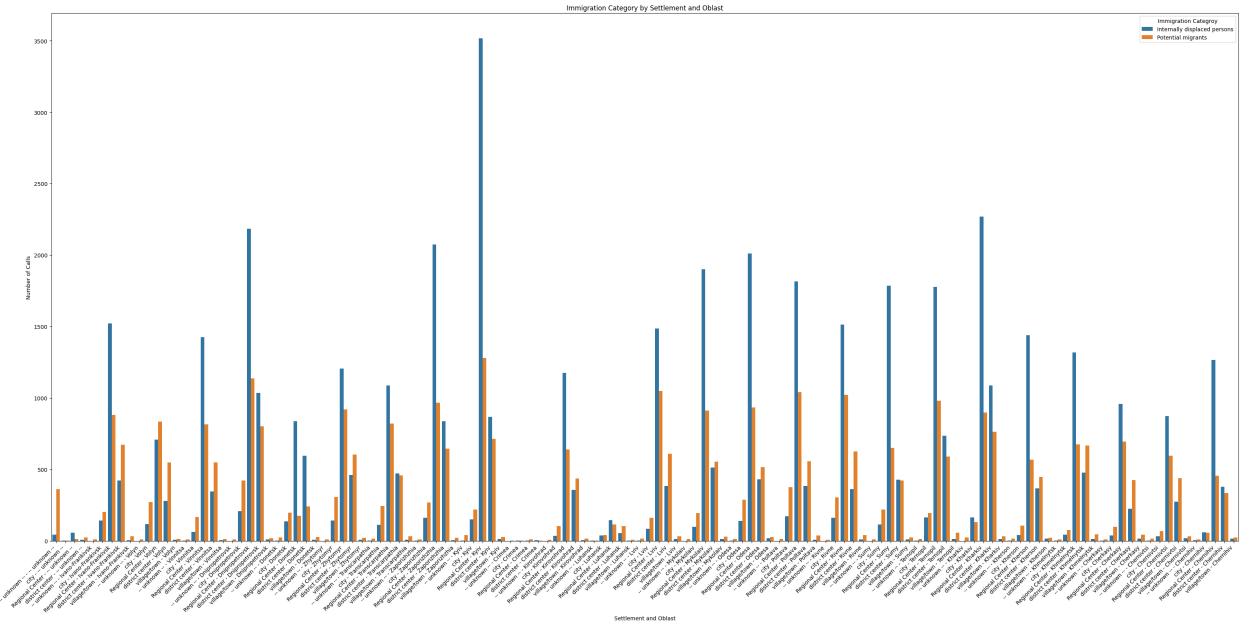


Understanding a figure that depicts the correlation among Settlement, Province, and Immigration Category (IDP versus Potential Migrants) is crucial in understanding the geographical dispersion and characteristics of distinct migrant groups. The spatial representation (see Figure 24, for a more legible view see Appendix E) of IDPs and Potential Migrants sheds light on regional migration trends and the diverse requirements of these populations in various locales.

By mapping the immigration categories across settlements and provinces (oblasts), decision-makers can identify areas with large numbers of IDPs or potential migrants. For example, based on the graphic

it appear as if Regional Centers and District Centers in Kharkiv, Kyiv and Dnipropetrovsk have greater numbers of IDPs compared to Potential Migrants, however, these regions have more numbers for both in comparison to the rest of the regions present. This information is important for several reasons. First of all, it helps understand the impact of migration and displacement on certain regions, which is essential for tailoring support services to unique needs of a community. For instance, areas with a high concentration of IDPs such as Kyiv and Kharkiv may need more emergency humanitarian help as well as support services for housing, food, and healthcare. Conversely, regions with a high concentration of prospective migrants could benefit from services focused on overseas employment opportunities, legal support, and migration information. Policymakers can use this spatial visualization to better allocate resources, ensuring that services and relief are distributed equitably and reach the most affected communities. Furthermore, combining immigration categories with settlement and provincial data analysis, reveals regional differences in movement and displacement patterns. These findings can prompt additional research into the underlying causes, which may include increased bombings, economic hardships, or violence.

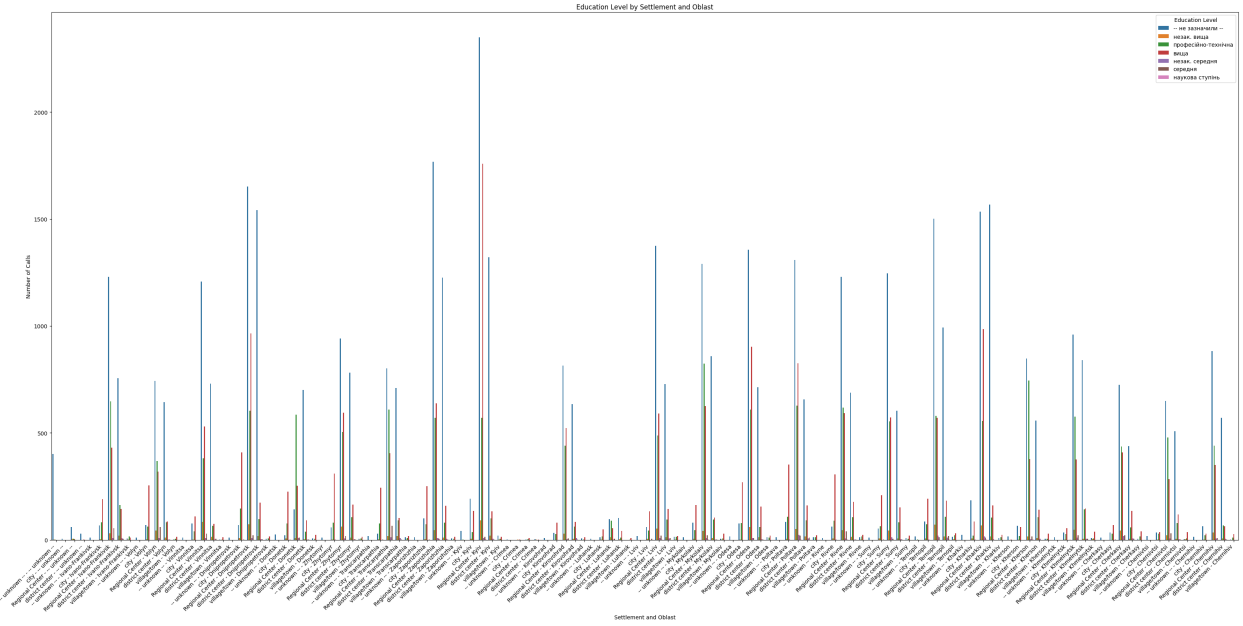
Figure 24: *Immigration Category by Settlement and Oblast (09/2021 - 09/2023)*



Understanding the geographic distribution of educational attainment and regional access to education need careful consideration of a visual that depicts the relationship between Settlement, Province, and Education Level (see Figure 25, for a more legible view see Appendix F). This analysis highlights locations that may need focused educational interventions and support, offering insightful information about how educational levels fluctuate among various settlements and provinces.

By mapping educational attainment across different geographic regions, researchers can identify patterns that indicate areas with lower levels of education. Policymakers and educational authorities need this information to develop strategies that address these disparities. For instance, areas with callers with lower levels of education may benefit from more funding educational infrastructure, and scholarship possibilities. This regional segmentation facilitates the identification of callers in particular regions that may lack adequate educational resources. It guarantees a more equitable distribution of educational opportunities by enabling the transfer of resources and support to the areas that require it the most. Additionally, since education is directly related to economic growth, employment possibilities, and general quality of life, knowing the relationship between settlement, province, and education level can provide insight into the larger socio-economic backdrop. An area of future study may involve an analysis with who is calling combined with educational levels to look at employment/education development program within specific regions.

Figure 25: *Education Level by Settlement and Oblast (09/2021 - 09/2023)*



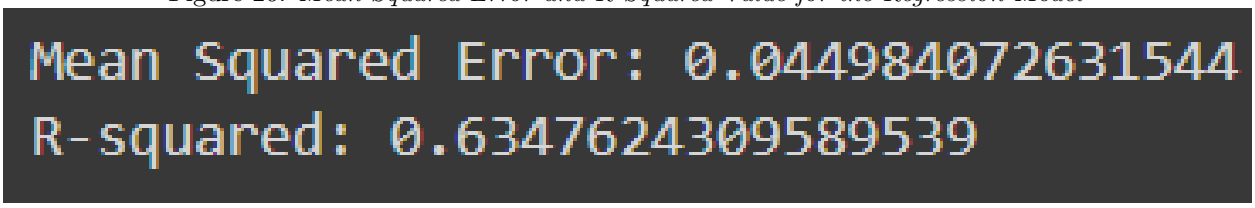
5.2.2 Regression Model Results

Finally, let's examine the results of the regression model. A regression is particularly valuable for analyzing "protection of the rights of Ukrainians abroad" for several reasons. First, regression analysis allows us to measure the correlation between this target variable and various other elements in the dataset. By identifying and quantifying the impact of multiple predictors, including categorical characteristics, economic indicators, and geographic variables, we can enhance our comprehension of the factors that motivate the defense of Ukrainian national rights overseas. This is especially crucial when it comes to international

relations and policy-making, where data-driven insights can guide practical measures to protect fundamental rights. Additionally, Regression models have predictive capabilities, enabling us to forecast the degree of protection under different conditions. This can aid in more efficient planning and resource allocation. The interpretability of regression models provides precise and useful insights into the most important components, directing future study and intervention efforts, particularly when feature importance analysis is included. Overall, regression modeling, facilities a comprehensive and detailed investigation of the factors influencing the defense of Ukrainians' rights abroad. This, in turn, helps inform better policy and decision-making, ensuring more effective and targeted strategies to safeguard these rights.

The regression model analysis yielded an R-squared value of 0.63 and a Mean Squared Error (MSE) of 0.044 as seen in the figure below (See FIGURE 26). These metrics indicate that the model's predictions are reasonably close to the actual values. A lower MSE signifies higher predictive accuracy . An R-squared value of approximately 0.63 suggests that the model's predictors can explain about 63% of the variance in the "protection of the rights of Ukrainians abroad" data. This significant percentage demonstrates that the model effectively explains the variation in the target variable and chosen features that relevant for this purpose. These findings underscore the model's effectiveness in identifying the critical elements affecting the defense of Ukrainians' rights abroad. They also point out areas in which the model could be further improved with additional data or more advanced methods. In sum, the model's performance indicates a strong correlation between the predictors and the target variable, offering stakeholders and policymakers important information for defending Ukrainians' rights abroad.

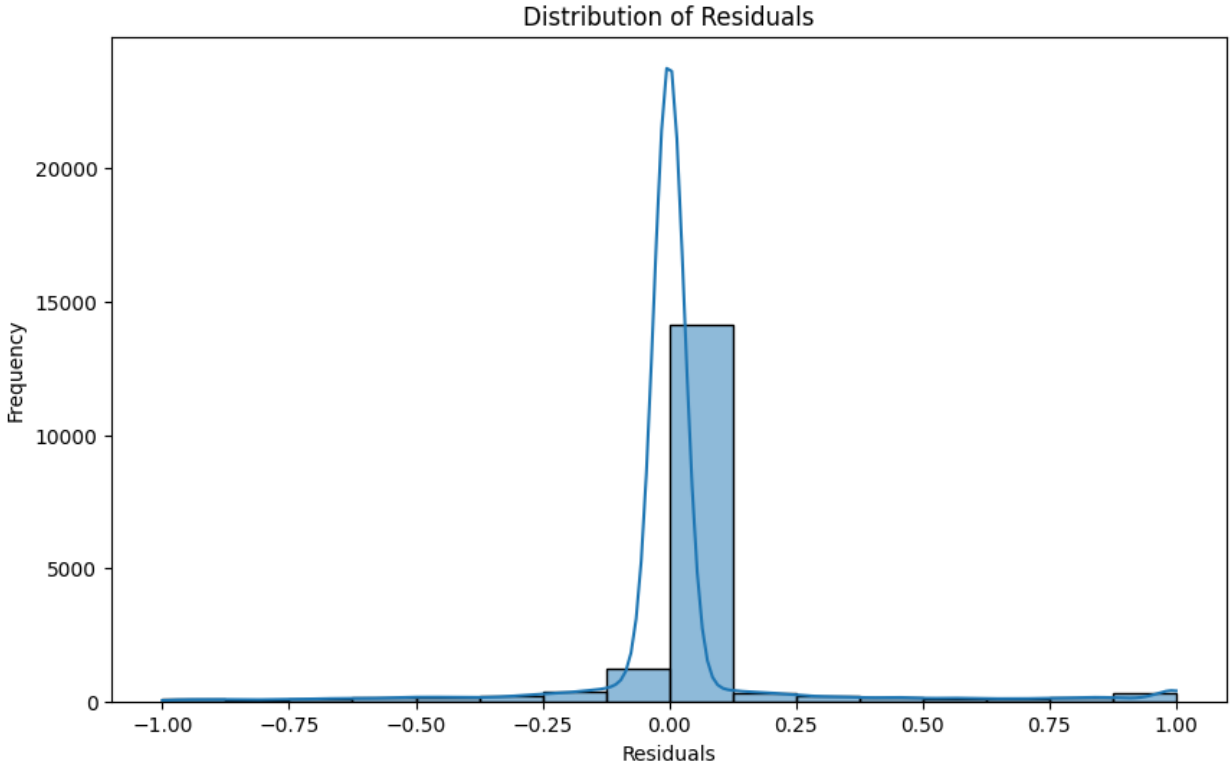
Figure 26: *Mean Squared Error and R Squared Value for the Regression Model*



Equally important, we consider a graphic of the distribution of residuals. FIGURE 27 below shows the residuals distribution from the regression model that predicted the "protection of the rights of Ukrainians abroad." The gap between the actual and projected values is are essential for assessing a regression model's performance. In the histogram, the residuals are highly centered around zero, indicating that the model's predictions are generally accurate and close to the actual values. The peak at zero, suggests a high frequency of low residuals, meaning that many predictions were almost correct. The clustering around zero, further indicates that the model does not consistently over-predict or under-predict the dependent variable.

On deeper analysis, there is a minor right skew in the distribution, with some residuals extending to the positive side. this skewness, suggests the model may under-predict the protection level in some cases, but these instances are few when compared to the central concentration around zero. The small number of residuals with high absolute values demonstrates the model's robustness and dependability. Overall, the residuals distribution aligns with the previous numerical evaluation results: the model accounts for a significant amount of the variation, as indicated by the low MSE and high R-squared value. This graphic confirms that the regression model is a useful tool for analysis and decision-making, accurately forecasting the preservation of Ukrainians' rights abroad.

Figure 27: *Residuals Distribution for the Regression Model*



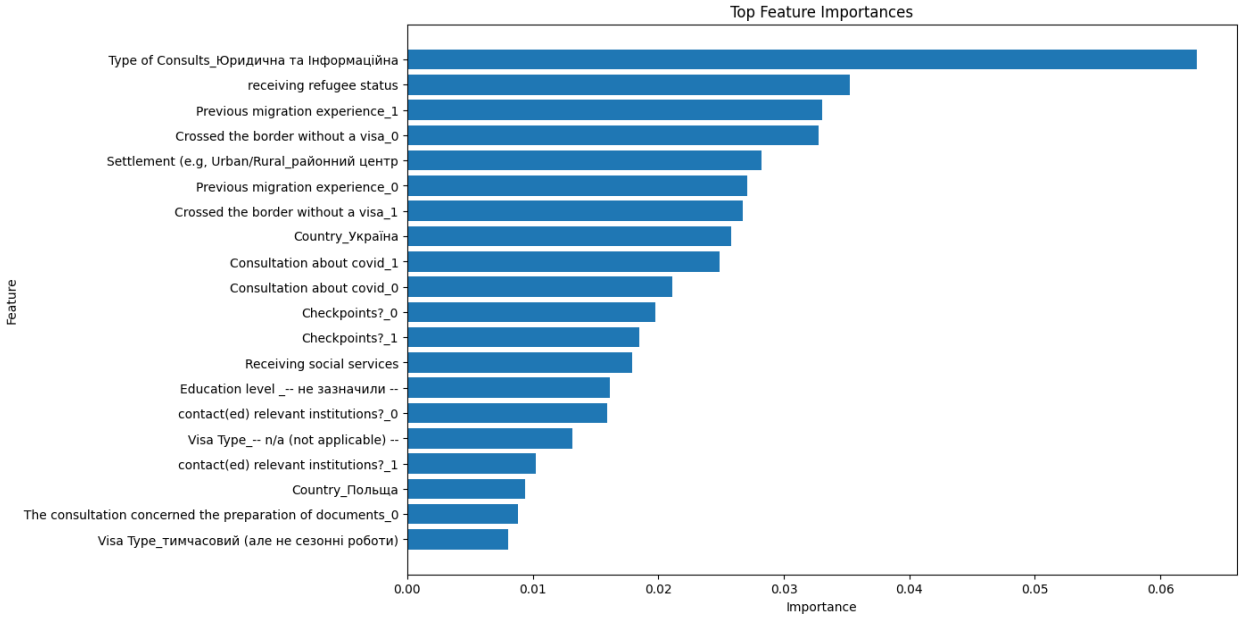
The regression model's feature importance chart provides a graphic depiction of the relative relevance of various predictors in understanding the "protection of the rights of Ukrainians abroad." This analysis offers insights into the elements that most significantly influence the target variable by highlighting the features have the most impact on the model's predictions.

The most prominent feature, with the highest relevance score, is **Type of Consults**. This suggests that legal and informative consultations play a crucial role in anticipating the defense of Ukrainians' rights abroad. The dominant importance of this feature implies that the kind of consultation is a key factor

influencing the model's prediction accuracy. Next in importance is **being granted refugee status**, which also demonstrates strong predictive power. The significant relevance score of this feature highlights the impact that formal refugee status has on the defense of rights for Ukrainians living abroad. Whether or not this status is granted appears to greatly influences the results of the model. Another notable predictor is **prior migration experience_1** . This suggests that previous migration experience positively impacts the protection of rights. The predominance of this trait implies that people with migration experience might be more knowledge or better able to protect their rights when moving to new locations.

In summary, the analysis of feature importance reveals that the type of consultation, receiving refugee status, and previous migration experience are the most critical factors in predicting the protection of Ukrainians' rights abroad. These insights can guide further research and policy-making by emphasizing the most influential factors identified in the model.

Figure 28: *Feature Importance Histogram for the Regression Model*



The regression model developed in this study is a valuable tool for forecasting how well Ukrainians' rights will be protected abroad. By identifying and quantifying the influence of several factors –including the type of consultations received, refugee status, and prior migration experience– this approach provides useful insights for policymakers and organizations working with Ukrainian migrants. For example, the importance of legal and informational consultations suggests that improving access to these services could significantly enhance the protection of rights for Ukrainians living abroad. Furthermore, understanding the impact of refugee status can help design more effective resources and support programs for people in this

situation. The model also highlights the significance of prior migration experience, indicating that educating and preparing individuals before migration could be beneficial. Overall, this regression model not only helps forecast outcomes but also provides information for targeted interventions and policy choices that improve the rights of Ukrainians living overseas.

6 Reflection

6.1 Discussion of design in the context of the project

The primary goal of this project's design was to implement robust industrial engineering methodologies to optimize and analyze complex datasets. By leveraging principles of process improvement and systems analysis, we aimed to create efficient and reliable frameworks for examining communications data from Telegram and the 527 hotline. This involved integrating advanced machine learning and data analysis techniques to derive meaningful insights, streamline data processing, and ultimately enhance the decision-making processes of organizations handling these sensitive datasets. Through meticulous design and rigorous validation, we ensured that our analytical procedures met the high standards required for dealing with such critical and confidential information.

1. **Defining the Problem:** The first and most important step was recognizing the need to derive useful insights from the Telegram conversations and 527 hotline data, with an emphasis on comprehending trends, salient characteristics, and hidden subjects associated with the conflict and humanitarian concerns.

2. **Requirement Analysis:** Requirements were defined based on the nature of the datasets and the specific objectives. for the 527 hotline data, the aim was to analyze the variables affecting Ukrainians' rights protection abroad. Another objective was to conduct a thorough EDA to understand how the invasion impacted the need for information and migration patterns. oFr the Telegram data, the objectives were to identify the latent issues and communication pattern.

3. **Solution Development:** The solutions included designing an EDA and regression model for the 527 helpline data, aiming to identify significant features impacting the target variable. For the Telegram data, the solution involved designing an LDA model to uncover latent topics, complemented by comprehensive EDA to understand communication trends.

4. **Implementation:** Using tools like Pandas for data manipulation, Scikit-learn for regression modeling, and LDAvis for topic modeling, the designed procedures were implemented in Python. Data preprocessing procedures, such as managing missing values, encoding categorical variables, and scaling numerical characteristics, were also a part of the implementation phase. In addition to the above, Dr. Dean will be presenting the results this summer (2024) to the IOM, if approved I would like to join the presentation as well.

5. **Validation:** Model performance criteria including Mean Squared Error (MSE) and R-squared for the regression model and the coherence and interpretability of topics for the LDA model were used to validate the efficacy of the designs. The results were then presented, and the insights gleaned from the models were validated using visualization tools.

by adhering to this methodical design process, The project effectively created analytical tools and models that offer insightful information about the datasets. This approach demonstrates the successful application of engineering design concepts to data science and machine learning initiatives.

6.2 Discussion of constraints considered in the design and broader impact

In this project, a multitude of constraints were considered to ensure ethical and effective outcomes. Data privacy and security were of utmost concern, which involves people providing private and sensitive information via the 527 hotline and Telegram messages. Strong security measures were therefore required to protect the data from cyberattacks and unlawful access, guarantee adherence to data protection laws, and uphold the confidence of the people who trusted us with their private information.

Ethical considerations were at the forefront, ensuring that the analysis was conducted with respect for the individuals represented in the data. As part of this, strict anonymization procedures were used to safeguard identities, and the insights were only used for the intended humanitarian ends. Some of these procedures involved undergoing CITI training so that I was aware of how to handle this information and how to handle any conflicts that may arise. Constraints related to health and safety were particularly important because the analysis's goal was to improve services and solutions for vulnerable people, improving their well-being in the process.

These limitations constrained the design options, resulting in the adoption of safe data handling procedures, ethical data analysis techniques, and models that could deliver useful insights without jeopardizing privacy. For example, secure access controls and encrypted data storage were implemented to mitigate cyber risks. Integrating ethical review procedures also helped to guarantee that the data was used responsibly.

These limitations significantly influenced the project's recommended measures, emphasizing the need for safe, morally righteous, and financially feasible alternatives that may improve the support networks for those who are experiencing a crisis. By enhancing data-driven decision-making in humanitarian interventions, the project has larger, global, economic, environmental, and social effects. Resources are allocated more effectively, policies are better informed, and affected individuals benefit from a more responsive and resilient

support system. The well-considered solutions are designed to enable organizations to offer prompt and efficient support, which will have a beneficial effect on the stability and well-being of society.

6.3 Discussion of your experience acquiring and applying new knowledge

First and foremost, learning about humanitarian efforts was something that was new to me throughout this project. To continually expand my understanding of humanitarianism my advisors Professor Konrad and Professor Dean provided me with many resources During our meetings, we would discuss these materials and address and any questions I had. This project spanned over a year year to be exact, and to stay on track I had to improve on my time-management skills to make sure that appropriate deliverables were on-time. To achieve this I had implemented the use of Gantt charts and set both firm and flexible deadlines for myself. In the end, this had helped me stay on task and meet deadlines for my project.

When implementing the project, there were several new concepts and techniques I had to learn. Although, I had previous experience with Machine Learning projects, I had never worked with topic modeling through LDA. To gain an understanding of this material, I had conversations with my Machine Learning Professor Sethi (Worcester Polytechnic Institute), and used simpler datasets to generate a simple model that I could refer to when working with this larger dataset. I also consulted literature on evaluation metrics for machine learning models. Throughout the project when there were things that I did not understand, it was valuable to have asked my advisor or Amir Jamali, a WPI School of Business Ph.D student heavily involved with this project. Their insights provided me with valuable insights on how to proceed. Otherwise, I would search the internet for reliable information and proceed accordingly. This combination of resources and support enabled me to effectively implement the project.

6.4 Discussion of teamwork in the project

Throughout the duration of this project, our team worked collaboratively and efficiently, fostering a leadership-driven and inclusive environment to achieve our objectives. We held Frequent meetings with our advisors and the larger project team throughout the year, ensuring continuous guidance and alignment with our goals. Regular interactions with Dr. Dean, who had valuable contacts with the 527 hotline and the International Organization for Migration (IOM), provided crucial insights into the specific needs and expectations of these organizations. These interactions allowed us to ask clarifying questions and tailor our analysis to produce meaningful and actionable insights.

Additionally, we held regular meetings with Professor Konrad and Amir to discuss the project's goals, track our progress, and receive constructive feedback on what I had achieved thus far. Initially, I collaborated closely with Leonardo Coelho, but due to personal reasons, I had to leave the project midway. Despite this, Professor Konrad and the team maintained a high level of coordination and inclusivity, ensuring a smooth transition and continued progress for his project. When I had rejoined the project team, we established clear goals, meticulously planned our work, and successfully met our objectives, demonstrating effective teamwork and a commitment to producing high-quality and important outcomes.

7 Conclusion

To uncover underlying themes and conversation topics in the dataset, we used a machine learning Latent Dirichlet Allocation (LDA) model and exploratory data analysis (EDA) on Telegram data. The EDA provided a comprehensive summary of the data, highlighting key trends, distributions, and correlations between different attributes. This initial analysis was essential for understanding the data's structure and substance, informing the next step of the modeling process.

The LDA model successfully identified distinct topics, revealing the most common themes that were discussed in the Telegram channels. These subjects ranged from community activities and social support to political debates and news updates. By examining the word distribution inside each topic, we gained significant insights about the main issues and interests of the Ukrainian refugees. This model effectively organizes and summarizes large amounts of text data, facilitating the identification of patterns and changes in public discourse.

The combination of EDA and LDA modeling has provided a comprehensive understanding of the Telegram data, offering valuable insights into the communication patterns and prevalent topics among users. This approach demonstrates the potential of machine learning techniques in analyzing social media data, which can be leveraged for various applications, including sentiment analysis, trend detection, and targeted communication strategies. Future research could build on this work by incorporating additional data sources and employing more advanced models to further enhance the accuracy and depth of the analysis.

In this study, to ascertain the variables impacting the defense of Ukrainians' rights overseas, we created a regression model and carried out a thorough exploratory data analysis (EDA) on the 527 helpline data. According to our model, a number of factors, including the kind of consultations (legal and informational in particular), refugee status, and prior migration experience, are significant factors that influence how well these rights are protected. A thorough summary of the data was given by the EDA, which also identified important correlations and trends that guided the regression model. The results imply that focused actions, such expanding access to particular kinds of consultations and aiding refugees and others with little prior experience migrating, could boost the efficacy of programs meant to defend Ukrainians' rights overseas. This approach offers data-driven insights to improve decision-making and maximize support services, making it an invaluable tool for organizations and governments. This foundation could be built upon in the future by incorporating more factors and investigating more sophisticated modeling strategies to improve forecasts

and suggestions. Regression modeling and EDA together have provided a thorough knowledge of the 527 helpline data, highlighting the significance of different aspects in forecasting the desired result. This strategy demonstrates how data-driven approaches can improve support services for vulnerable groups and inform policy decisions. This model could be improved in the future by including more variables and investigating more sophisticated regression techniques to further improve predicted performance.

Expanding upon the knowledge acquired from the data analysis of the 527 hotline, subsequent studies may investigate several directions to further our comprehension and amplify the influence of this endeavor. An important area of research is expanding the dataset to include more characteristics, like socioeconomic status, regional variations, and more precise demographic data, that capture the complex experiences of people seeking assistance. This could increase our models' forecast accuracy and provide us a more thorough grasp of the variables affecting the preservation of rights overseas. Additionally, the prediction power of the model may be further increased by utilizing more sophisticated machine learning approaches, such as ensemble methods or deep learning. This methodological paradigm could also be used to examine other geopolitical contexts, such the conflict between Israel and Palestine. Researchers can determine the most important problems that impacted communities confront, assess the efficacy of support services, and provide guidance for policy responses by examining hotline data or comparable datasets from conflict zones. Examining the specific ways in which various forms of support—legal, psychological, or material—affect people's rights and well-being in these areas could yield insightful information. In addition to highlighting similarities and variations in the difficulties encountered and the effectiveness of various intervention measures, comparative studies in war zones may also advance our understanding of humanitarian assistance and conflict resolution on a global scale.

A Appendix A: LDA Code

```
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer, PorterStemmer
from nltk.tokenize import word_tokenize
import gensim
from gensim.models import Phrases
import pyLDAvis
import pyLDAvis.gensim
import pandas as pd

# Function to preprocess text and include bi-grams and tri-grams
def preprocess_text(text):
    nltk.download('stopwords')
    nltk.download('punkt')
    nltk.download('wordnet')
    stop = set(stopwords.words('english'))

    stemmer = PorterStemmer()
    lemmatizer = WordNetLemmatizer()

    # Tokenize and clean the text
    def tokenize_and_clean(news):
        words = [w for w in word_tokenize(news.lower()) if w.isalpha() and w not in stop]
        words = [lemmatizer.lemmatize(w) for w in words if len(w) > 2]
        return words

    # Create the initial corpus
    corpus = [tokenize_and_clean(doc) for doc in text]

    # Generate bi-grams and tri-grams
```

```
    phrases = Phrases(corpus, min_count=5, threshold=100)
    bigram = Phrases(phrases[corpus], threshold=100)
    trigram = Phrases(bigram[phrases[corpus]], threshold=100)

    # Apply the phrases to the corpus
    corpus = [trigram[bigram[phrases[doc]]] for doc in corpus]

    return corpus

# Function to get LDA objects
def get_lda_objects(text):
    corpus = preprocess_text(text)

    dic = gensim.corpora.Dictionary(corpus)
    bow_corpus = [dic.doc2bow(doc) for doc in corpus]

    lda_model= gensim.models.LdaMulticore(bow_corpus, num_topics=8, id2word=dic, passes=10, workers=8)
    return lda_model, bow_corpus, dic

# Function to create the visualization
def plot_lda_vis(lda_model, bow_corpus, dic):
    pyLDAvis.enable_notebook()
    vis = pyLDAvis.gensim.prepare(lda_model, bow_corpus, dic)
    return vis

# Assuming 'combined_df' is your DataFrame and 'message_english_cleared' is the text column
lda_model, bow_corpus, dic = get_lda_objects(combined_df['message_english_cleared'])

# Generating the visualization
plot_lda_vis(lda_model, bow_corpus, dic)
```

B Appendix B: Regression Model Code

```
from sklearn.decomposition import PCA  # Import PCA

# Define the target variable and features
target = 'protection of the rights of Ukrainians abroad'
X = df.drop(columns=[target])
y = df[target]

# Convert all categorical columns to strings
categorical_cols = X.select_dtypes(include=['object', 'category']).columns
for col in categorical_cols:
    X[col] = X[col].astype(str)

# Identify numerical columns
numerical_cols = X.select_dtypes(include=['int64', 'float64']).columns

# Preprocess categorical features
categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='constant', fill_value='missing')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
])

# Preprocess numerical features with PCA for dimensionality reduction
numerical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='mean')),
    ('scaler', StandardScaler()),
    ('pca', PCA())  # No need to specify n_components here
])

preprocessor = ColumnTransformer(
    transformers=[
```

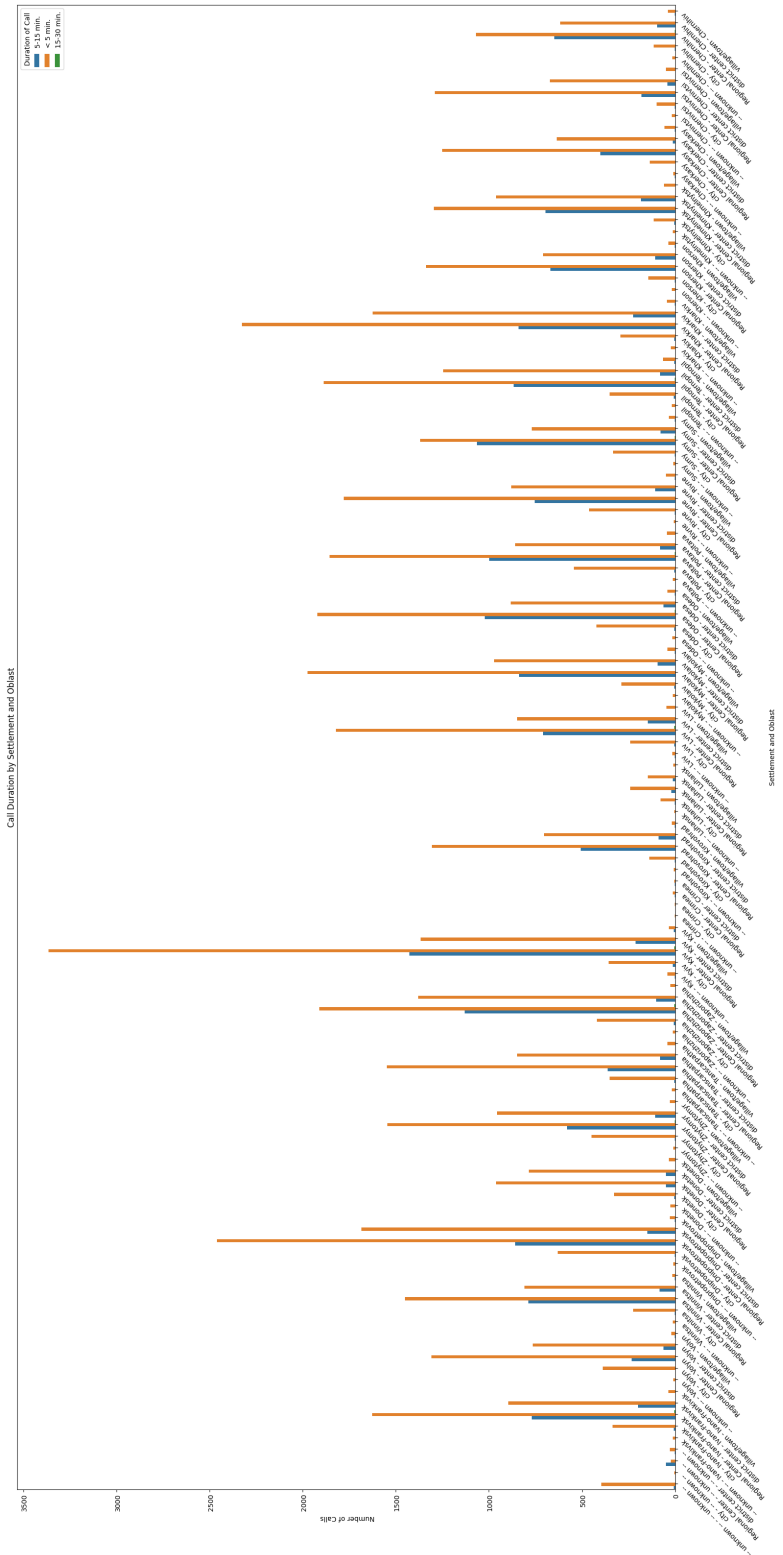
```
        ('num', numerical_transformer, numerical_cols),
        ('cat', categorical_transformer, categorical_cols) ])

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Define the model
model = RandomForestRegressor(n_estimators=100, random_state=42)

# Create a pipeline that includes preprocessing, PCA, and the model
clf = Pipeline(steps=[('preprocessor', preprocessor), ('regressor', model)])

# Train the model
clf.fit(X_train, y_train)

# Make predictions
y_pred = clf.predict(X_test)

# Convert 'missing' to NaN and handle missing values
y_test = y_test.replace('missing', float('nan')).astype('float')
y_pred = pd.Series(y_pred).replace('missing', float('nan')).astype('float')

# Impute NaN values in y_test and y_pred
imputer = SimpleImputer(strategy='mean')
y_test_imputed = imputer.fit_transform(y_test.values.reshape(-1, 1))
y_pred_imputed = imputer.transform(y_pred.values.reshape(-1, 1))

# Evaluate the model
mse = mean_squared_error(y_test_imputed, y_pred_imputed)
r2 = r2_score(y_test_imputed, y_pred_imputed)

print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')
```

C Appendix C: Visualizing Regression Model Code

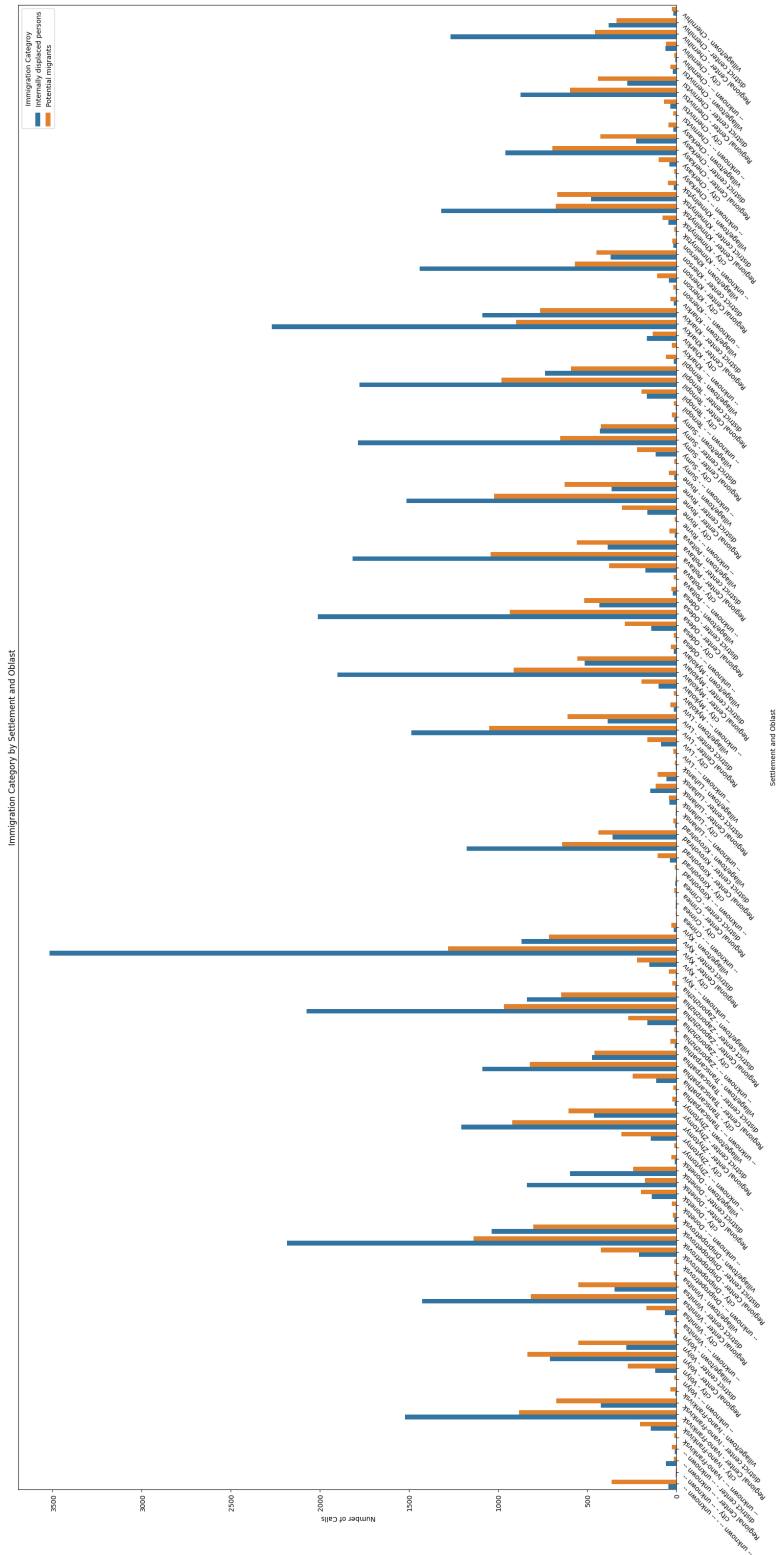
```
# Histogram of residuals
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.histplot(residuals, kde=True)
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.title('Distribution of Residuals')
plt.show()



# Get feature importances
importances = model.feature_importances_

# Get the feature names
onehot_columns = clf.named_steps['preprocessor'].named_transformers_['cat']['onehot'].get_feature_names
all_feature_names = np.concatenate([numerical_cols, onehot_columns])

# Create a DataFrame for feature importances
feature_importances = pd.DataFrame({
    'Feature': all_feature_names,
    'Importance': importances
})

# Sort the features by importance
feature_importances = feature_importances.sort_values(by='Importance', ascending=False)

# Select top N features
top_N = 20  # Adjust this number to display more or fewer features
top_features = feature_importances.head(top_N)
```

```
# Print the most important features
print(top_features)

# Plot the feature importances
plt.figure(figsize=(12, 8))
plt.barh(top_features['Feature'], top_features['Importance'])
plt.xlabel('Importance')
plt.ylabel('Feature')
plt.title('Top Feature Importances')
plt.gca().invert_yaxis()
plt.show()
```

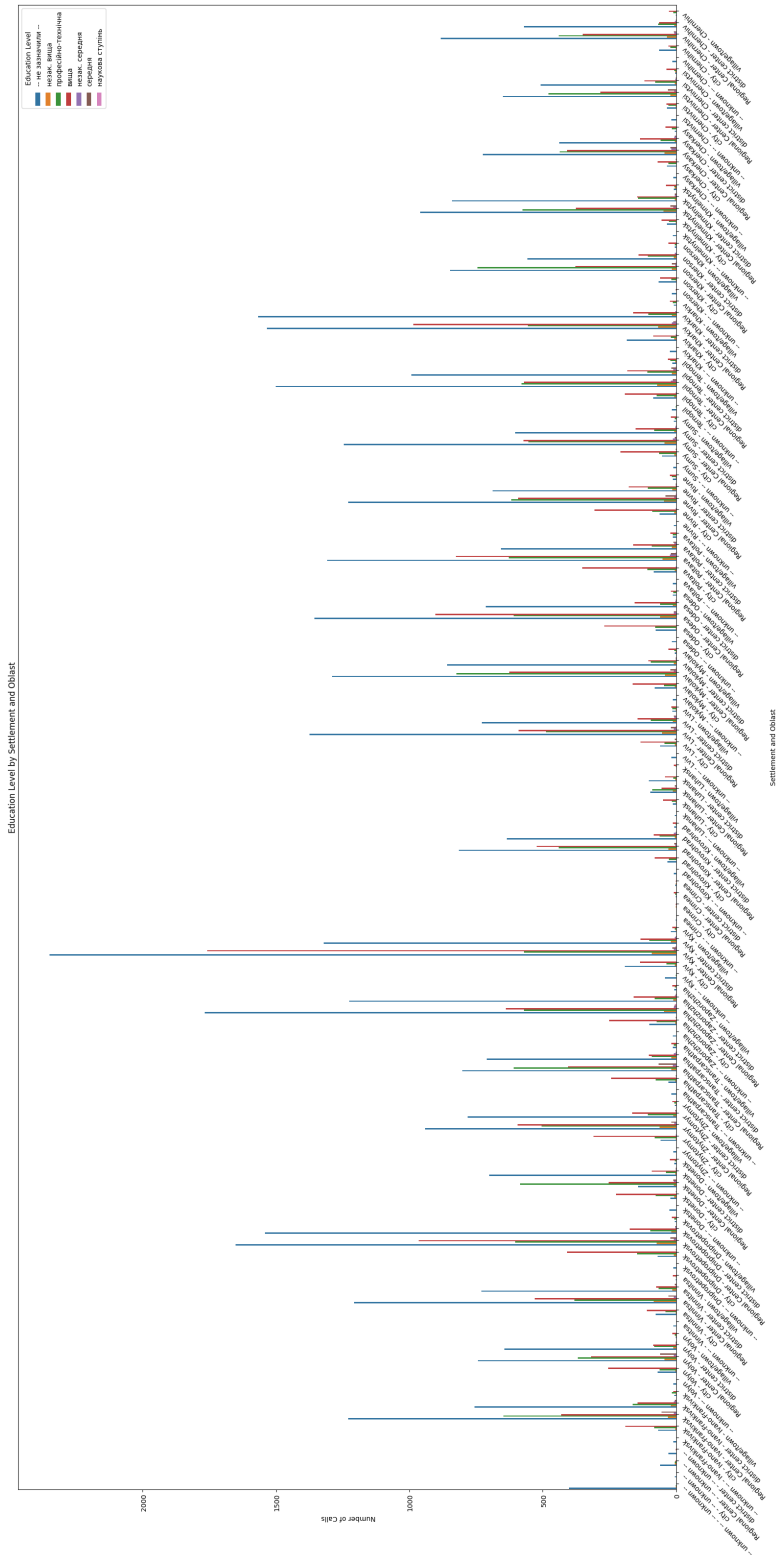
D Appendix D: Figure 23 Enlarged



E Appendix E: Figure 24 Enlarged



F Appendix F: Figure 25 Enlarged



References

Alvarez, H. and Serrato, M. (2013). Social network analysis for humanitarian logistics operations in latin america. *IIE Annual Conference and Expo 2013*.

Amnesty International, A. I. (2022). Ukraine: Humanitarian corridors for civilians fleeing Russian attacks must provide safety – new testimonies.

Blei, D., Ng, A., and Jordan, M. (2001). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:601–608.

Cockbain, E. and Tompson, L. (2024). The role of helplines in the anti-trafficking space: examining contacts to a major 'modern slavery' hotline. *CrimRxiv*. https://www.crimrxiv.com/pub/e1y6588f.

Daddoust, L., Asgary, A., McBey, K. J., Elliott, S., and Normand, A. (2021). Spontaneous volunteer coordination during disasters and emergencies: Opportunities, challenges, and risks. *International Journal of Disaster Risk Reduction*, 65:102546.

Dunn, C. and Kaliszewska, E. (2023). Distributed humanitarianism. *American Ethnologist*, 50(1):19–29.

Ferris, E. (2018). When refugee displacement drags on, is self-reliance the answer? Accessed: 2024-05-30.

Haq, Sana Noor, N. E. (2023). 'Complete paralysis:' Palestinian medics say disaster awaits Gaza as Israel pounds enclave with airstrikes.

Hjalmar Bang Carlsen, T. G. and Toubøl, J. (2023). Ukrainian refugee solidarity mobilization online. *European Societies*, 0(0):1–12.

ICRC, I. (2022). How humanitarian corridors work to help people in conflict zones.

Juggins, S. and Telford, R. J. (2012). *Exploratory Data Analysis and Data Display*, pages 123–141. Springer Netherlands, Dordrecht.

Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Kelechava, M. (2020). Using lda topic models as a classification model input.

Kershner, I. and Shaar-Yashuv, A. (2023). 'United Because of This Disaster': Israelis Rush to Volunteer After Hamas Attacks. *The New York Times*.

Lipka, M. (2022). Attitudes on taking in refugees vary by party, race and ethnicity.

Munková, D., Munk, M., and Vozár, M. (2014). Influence of stop-words removal on sequence patterns identification within comparable corpora. pages 67–76.

Poushter, J. (2016). European opinions of the refugee crisis in 5 charts.

Reif, E., Qian, C., Wexler, J., and Kahng, M. (2024). Automatic histograms: Leveraging language models for text dataset exploration.

UNHCR (2017). What is a refugee? Accessed: 2024-05-30.

UNHCR (2018). Basic needs approach in the refugee response. Accessed: 2024-05-30.

UNHCR (2020). Refugees in america. Accessed: 2024-05-30.

UNHCR (2023). Ukraine refugee crisis. Accessed: 2024-05-30.

Ye, J., Jindal, N., Pierri, F., and Luceri, L. (2023). *Online Networks of Support in Distressed Environments: Solidarity and Mobilization during the Russian Invasion of Ukraine*. ICWSM.