10-9-2023

Analyzing the Needs of Ukrainian Refugees on Telegram in Real-Time: A Machine Learning Approach

Raphael Reimann
Hasso-Plattner-Institute, Potsdam, Germany, raphael.reimann@student.hpi.uni-potsdam.de

Matthew Caron
Paderborn University, Paderborn, Germany, matthew.caron@uni-paderborn.de

Analyzing the Needs of Ukrainian Refugees on Telegram in Real-Time: A Machine Learning Approach

Research Paper

Raphael Reimann[1] and Matthew Caron[2]

[1] Paderborn University, Paderborn, Germany
raphael.reimann@student.hpi.uni-potsdam.de
[2] Paderborn University, Paderborn, Germany
matthew.caron@uni-paderborn.de

Abstract. The humanitarian crisis resulting from the Russian invasion of Ukraine has led to millions of displaced individuals across Europe. Addressing the evolving needs of these refugees is crucial for hosting countries and humanitarian organizations. This study leverages social media analytics to supplement traditional surveys, providing real-time insights into refugee needs by analyzing over two million messages from Telegram, a vital platform for Ukrainian refugees in Germany. We employ Natural Language Processing techniques, including language identification, sentiment analysis, and topic modeling, to identify well-defined topic clusters such as housing, financial and legal assistance, language courses, job market access, and medical needs. Our findings also reveal changes in topic occurrence and nature over time. To support practitioners, we introduce an interactive web-based dashboard for continuous analysis of refugee needs.

Keywords: Telegram Analysis, Machine Learning, Topic Modeling, Evidence-based decision-making, Humanitarian assistance.

1 Introduction

The Russian Federation's large-scale invasion of Ukrainian territories, initiated on February 24, 2022, has profoundly affected European migration patterns and the integration of displaced persons (Brücker et al. 2022). In just five days following the beginning of the invasion, over 500,000 Ukrainians were forced to leave their country. The European Union (EU) visa exemption for Ukrainian citizens has significantly impacted the migration potential into the EU (Trauner & Valodskaite 2022). Since the invasion began, over 5.5 million Ukrainian refugees have been recorded across Europe, of that nearly 900,000 in Germany alone (UNHCR 2022*b*). As of June 2022, 2,000 to 3,000 new Ukrainian refugees are registered in Germany daily, marking the fastest increase in refugee migration since World War II (Brücker 2022). The refugee demographic is skewed towards children, with around 50% of Ukrainian refugees coming to Germany, while most adult refugees are women (Brücker et al. 2022). The uncertainty surrounding the invasions trajectory affects the integration of refugees into the job market and German society, as their acceptance of language and integration courses is dependent on the conflict's progression (Brücker et al. 2022).

18th International Conference on Wirtschaftsinformatik
September 2023, Paderborn, Germany

Host countries like Germany now have to provide for incoming refugees from Ukraine. To coordinate such a post-disaster situation, it is essential to understand what resources are needed (Basu et al. 2019) and to comprehend the needs and requirements of incoming refugees through direct communication (Khatua & Nejdl 2021). Current surveys on the needs of Ukrainian refugees are limited to a specific timeframe and require expert interviewers (UN Women 2022, Bundesministerium des Innern und für die Heimat 2022, Pötzschke et al. 2022), whereas analyzing digital data sources can provide continuous monitoring of refugees' needs (Adema & Maitreyee 2022). This potential is currently underutilized with a focus on public perception through English Twitter streams and not the voices of Ukrainian refugees (Chen & Ferrara 2022, Zhu et al. 2022, Pohl et al. 2022).

Since the beginning of the ongoing crisis, big data analytics approaches have been used to analyze forced migration flows from the Russian invasion of Ukraine (Pötzschke et al. 2022). For instance, the World Health Organization recommends using social media data analytics tools to detect public health threats in countries hosting displaced Ukrainian refugees (World Health Organization et al. 2022). In fact, internet measurements have served as a proxy to examine refugee flows from Ukraine into neighboring countries (Mizrahi & Yallouz 2022). Other approaches used to investigate the crisis in Ukraine include analyzing Google Trends queries, sentiment analysis of tweets, and social media activity assessments (Jurić 2022, Adema & Maitreyee 2022, Caprolu et al. 2022, Park et al. 2022).

Nevertheless, a research gap persists in understanding the needs of Ukrainian refugees without imposing pre-established categories, which is vital for informing policy decisions by authorities and humanitarian organizations (Adema & Maitreyee 2022). This motivates the investigation of the following research question:

How can machine learning methods help
identify and analyze the needs of Ukrainian refugees?

To address this, we have identified more than 130 Telegram discussion channels used by Ukrainian refugees in Germany, collected over 2 million Telegram messages, and analyzed the data using machine learning approaches such as language identification, sentiment classification, and topic modeling. Our study compared two topic modeling approaches, LDA and transformer-based BERTopic, to identify topic clusters statically and over time. Our analysis revealed distinct clusters of needs for Ukrainian refugees in Germany with key areas of concern, including housing, financial assistance, healthcare, language, integration courses, legal assistance, translation support, and safe transportation opportunities across borders. These findings have been structured into a conceptual framework that can guide authorities and humanitarian organizations in addressing incoming refugees' immediate and long-term needs. This framework not only aids in understanding the current situation but also provides a versatile tool that can be adapted to different crisis scenarios. We hope this model will enhance policy-making and resource allocation processes, contributing to a more effective and tailored response to refugee crises.

2 Related Work

2.1 Understanding the Needs of Ukrainian Refugees

The United Nations High Commissioner for Human Rights (UNHCR) conducted interviews with Ukrainian refugees in border locations, transit centers, and assistance points in Belarus, Bulgaria, Hungary, Republic of Moldova, Poland, Romania, and Slovakia (UNHCR 2022*a*), identifying urgent needs including cash, employment, accommodation, material assistance, healthcare, food, education, family reunification, legal advice, information about services, transportation, psychological support, and a medium to contact family. Notably, social media, especially Telegram, was identified as the preferred information channel for refugees, second only to Facebook. Similarly, a survey conducted by the German Federal Ministry of the Interior and Home Affairs found that financial aid, medical care, and assistance with administrative procedures are the most essential needs of refugees (Bundesministerium des Innern und für die Heimat 2022). Yet, only 19% of respondents stated that they could work in Germany, suggesting the language barrier is a significant obstacle for incoming refugees. Meanwhile, a study analyzing the online search behavior of Ukrainian refugees using Google Trends also recognized a shift in interest towards evacuation-related topics as the invasion intensified, but the study was limited by its reliance on predefined keywords (Adema & Maitreyee 2022).

2.2 Social Media in Crisis Situations

Social media platforms have played an integral role during political crisis events such as the Arab Spring and the Euromaidan protests in Ukraine (Sabatovych 2019). Various studies have used topic modeling to analyze Twitter messages from the first two weeks of the Russian invasion of Ukraine (Skwarek et al. 2022), explored digital media usage by refugees migrating to Europe (Gillespie et al. 2016), or used keyword searches on Google Trends to analyze refugee flows from Ukraine to the EU (Jurić 2022). Social media analytics have also guided public health policies during disease outbreaks (Yigitcanlar et al. 2020). In refugee situations, prior research has provided input on using social media to monitor refugee migration decisions during the Syrian war to support policymakers (Walk et al. 2023). Such research underscores the potential of social media as a critical tool in crisis management (Imran et al. 2015, Li et al. 2017, Nazer et al. 2017). While prior research has focused primarily on platforms like Twitter and Google Trends, our study takes a novel approach by focusing on Telegram, which offers a more intimate and continuous form of communication and may provide more in-depth insights into the evolving needs of refugees.

2.3 Telegram

Telegram enables private chats, public channels, and groups, with this study focusing exclusively on publicly accessible channels. Users and channels can forward messages to other users, groups, and channels. In this study, we will not distinguish between channels and groups since this distinction is irrelevant to our findings. Unlike other social media platforms, Telegram's private and unsearchable nature encourages open and honest discussions among refugees, offering a safe space for sharing needs and challenges. This unique context can provide valuable insights unavailable on other social platforms.

Telegram has been the subject of various research studies, including analyses of its structural and topical features (Dargahi Nobari et al. 2017), the effects of government bans on the platform (Akbari & Gabdulhakov 2019), its role in protest mobilization (Urman et al. 2021), and discussions of public opinion on topics such as COVID-19 (Ng & Loke 2020). Telegram has also been linked to extremist activities and political messaging, with some studies pointing to its potential as a platform for hate speech and disinformation (Urman & Katz 2020, Wich et al. 2021, Knuutila et al. 2020). Despite its use and potential impacts, Telegram has received relatively little attention compared to other messaging platforms in academic research (Hoseini et al. 2020).

2.4 Natural Language Processing

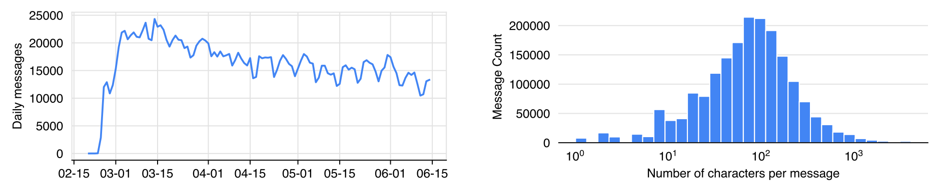
Natural Language Processing (NLP) is the research field focused on analyzing text data to generate insights into human relationships, needs, and desires. Recent advancements, such as transfer learning and the Transformer architecture (Vaswani et al. 2017), have improved the analysis of unstructured data produced on social media, which can contain short texts, emojis, or internet slang (Alyafeai et al. 2020). An example of how an NLP model can be applied is presented by Barbieri et al. (2022), who fine-tuned a XLM-R language model (Conneau et al. 2020) on 198 million tweets retrieved from Twitter, training one additional classification layer on the task of sentiment analysis. They found significant potential regarding zero-shot performance, meaning the model also performs on previously unseen languages. The XLM-Twitter model used in our study builds on the work of Barbieri et al. and is particularly suited for multilingual sentiment analysis on social media data.

3 Data

In this work, we analyze Telegram channels dedicated to Ukrainian refugees migrating to Germany, including city/demographic-specific, general refugee, German citizen, and refugee-support organizations channels. We initially curated a list of 30 channels with the assistance of domain experts and other social media platforms. We then used a snowball method to identify additional relevant channels through the Telegram API that fit our criteria by tracing forwarded messages. This iterative process continued until there were no new channels to discover. We finally manually removed irrelevant channels to our study, collected all messages of the remaining channels, and stored them in a relational database.

We focus on analyzing the needs of Ukrainian refugees in Germany following the invasion that began on February 24, 2022, focusing on their evolution over time. Moreover, informed by our findings, we also aim to find a real-time method for analyzing new incoming messages. To achieve this, we partitioned our dataset into train and test sets based on the timestamp of the messages in order to prevent data leakage. Our dataset is separated at an arbitrary cut-off date of 15.06.2022, resulting in a train set containing approximately 75% of all messages.

Our dataset of 1,859,135 messages was collected from 136 channels, most of which were created after the invasion of Ukrainian territories in February 2022. Relevant channels include *Help for Ukrainians in Germany* (translated from Ukrainian) or *Berlin helps Ukrainians*. The most popular channels have over 1,000 members and a median of 2,089.5 members. Channels have an average of 13,978.46 messages, with a median of 720. In Fig 1a, we show the message volume across all channels over time. We find an increase in message volume until early March 2022 and an overall decline until our cut-off date. The data also reveals weekly patterns, with a lower message volume on the weekends. The median message length is 76 characters, with the distribution of message length displayed in Fig 1b, indicating that messages tend to be short.



(a) Message volume over time **(b)** Distribution of message character count

Figure 1. Distribution of daily message volume and character count per message

4 Methods

In our analysis of the collected messages, we applied several methods. Initially, we identified the language of each message, motivated by potential differences in discussed topics across languages and the possibility of inferring the sender's primary language. Next, we analyzed the sentiment of each message to determine whether the discussed topics were positive or negative. This helped indicate whether refugee needs were being met. Lastly, we applied topic modeling to reveal different clusters of topics within the messages that could hint at different refugee needs.

4.1 Language Identification

Automatic Language Identification is the task of identifying the language of a document (Baldwin & Lui 2010). We used the pre-trained FastText model *lid.176.bin* for language identification, provided by Joulin et al. (2016). Through manual inspection of the dataset, we identified Russian, Ukrainian, English, and German as the majority of languages used in our dataset. Messages that could not be classified into one of those four languages, for instance, those containing only URLs, emojis, or ambiguous words, were labeled as *Other*. As the FastText model returns a language label and a probability for a given input string, we labeled all messages with an identified language probability of < 0.6 or a label other than *Russian*, *Ukrainian*, *English*, or *German* as *Other*.

4.2 Sentiment Analysis

While sentiment analysis is well-researched for English and other high-resource languages, it is not well-studied for Russian content (Smetanin 2020) and even less so for Ukrainian language cases (Bergsma et al. 2012). Considering that our corpus contains several languages, we opted for a pre-trained model for sentiment analysis. This decision was informed by the multilingual nature of our data and the lack of annotated data regarding the messages' sentiment. We utilized the XLM-T Transformer model to predict Telegram message sentiment, addressing the scarcity of pre-trained multilingual models for social media (Barbieri et al. 2022). Trained on Twitter data, the XLM-T model is sensitive to domain-specific language characteristics, including misspellings, slang, vulgarisms, and emojis. While the languages used on Telegram and Twitter may vary, some characteristics overlap, including the relatively short message lengths (280 characters on Twitter, 4096 on Telegram), comparable use of emojis and slang, and frequent grammatical errors and misspellings. Less than 10% of the messages in our Telegram training set exceed Twitter's limit, suggesting that XLM-T should perform reasonably well on our Telegram data.

4.3 Topic Modeling

Topic modeling of social media content can be challenging due to its short and unstructured nature, which often includes compound words, acronyms, and grammatical errors. Previous research used Latent Dirichlet Allocation (LDA) as a standard approach, despite its limitations with noisy or sparse text and neglected evaluations (Egger & Yu 2022, 2021). To widen the scope of topic modeling, we evaluate the performance of two approaches: LDA and BERTopic, which utilizes performant language models. We compare the strengths and weaknesses of LDA as a well-established method and BERTopic as a modern alternative (Grootendorst 2022).

Latent Dirichlet Allocation. Latent Dirichlet Allocation (LDA) is one of the most widely used methods for topic modeling (Blei et al. 2003). It is a generative probabilistic model that allocates a mixture of topics for each document and a distribution of words for each topic. Standard processing steps are usually followed when training LDA topic models, which include N-gram tokenization, removal of stop words, and lemmatization (Debortoli et al. 2016). We trained LDA models to find the optimal number of topics per language, using coherence and perplexity as evaluation metrics. A good trade-off between high coherence and low perplexity was achieved by choosing the number of topics for each LDA model as 18 for the Russian language, 9 for the Ukrainian language, 20 for the English language, and 20 for the German language.

BERTopic. BERTopic is a topic modeling technique that extracts coherent topic representation by developing a class-based variation of TF-IDF (Grootendorst 2022). It uses pre-trained transformer-based language models to create embeddings for given documents. Dimensionality reduction is applied to those embeddings, after which they are clustered. BERTopic consistently outperforms traditional techniques such as LDA in standard metrics used to evaluate topic models, such as topic coherence and topic diversity, which indicate the association of documents within a topic and how varied topics are. For our analysis, we removed punctuation, URLs, and linebreak characters from the messages and trained a BERTopic model using the pre-trained *paraphrase-multilingual-MiniLM-L12-v2* Sentence Transformer model (Reimers & Gurevych 2019). Due to memory constraints, the BERTopic model was trained on a 30% random subset of the training set.

5 Results

In this section, we present the results of our language identification and sentiment analysis methods. The language identification model achieved a weighted F1-score of 0.951, indicating good performance in predicting the language of messages in our dataset. Most messages were identified as Russian, followed by Ukrainian, English, and German. About 10% of all remaining messages were classified as other, containing very short texts, URL links, or emojis.

Figure 2 shows the percentage of messages by language over time, with English and German messages decreasing over time, while Russian and Ukrainian messages remained stable. Based on manual inspection, the decrease in English and German messages can be attributed to declining support from local German individuals and the professionalization of refugee organizations using the native languages of refugees, mainly Ukrainian and Russian.

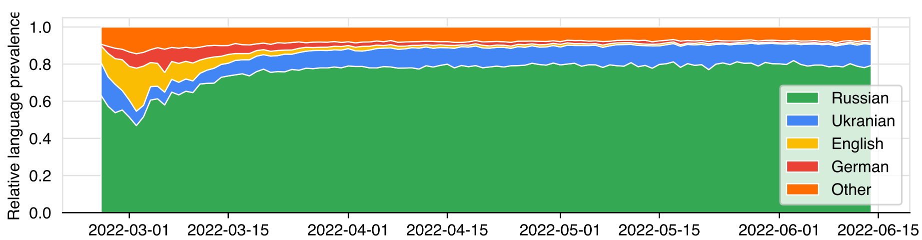


Figure 2. Language of messages over time

Sentiment Analysis. The sentiment classification model achieved a weighted F1-score of 0.939, suggesting good performance in predicting our dataset's messages' sentiment. Most messages were classified as neutral (80%), followed by negative (11%) and positive (9%). Going forward, we calculate the average sentiment as the simple sum over the sentiment values since we have classified negative, neutral, and positive messages as -1, 0, and 1, respectively.

Russian LDA. The Russian LDA model was trained on 18 topics with a corpus of 1,403,479 documents. The topics generated by the model are specific to refugee needs. These topics include questions about free services for Ukrainian refugees, access to doctors, housing, schooling for children and adults, work-related matters and assistance, financial assistance for refugees from the German government, transportation across borders, and housing searches. However, the model also generated general topics that do not represent refugee needs, such as people's interaction with one another and people asking questions and thanking others for answering their questions.

Ukrainian LDA. The Ukrainian LDA model was trained on ten topics with a corpus of 179,797 documents. The resulting model contains specific topics such as the search for housing or what documents are needed when crossing a border. Two additional topic clusters describe the need for information on how German language qualifications and courses can be found and general questions about the German job market, such as work hours and compensation. We also find a cluster of messages about moral support for the Ukrainian military and national pride with common expressions such as "glory_ukraine" (translated from Ukrainian). This model also produces topic clusters containing general dialogue content unrelated to refugee needs.

English LDA. The English LDA model is trained on 20 topics with a corpus of 61,779 documents, covering various aspects of refugee-related messages. One topic cluster focuses on language qualification for refugees in the context of work and university. Another topic cluster is related to travel, specifically by train from Ukraine through Poland and finally to Berlin in Germany, which is a common route taken by Ukrainian refugees arriving in Germany. Accommodation and support are the main themes in two topic clusters, with refugees searching for places to stay or offering translation services. Medical attention is the primary theme in another topic cluster. Other topics clusters cover calls by refugee organizations seeking volunteers and donations of humanitarian goods. Finally, several topics are unrelated to refugee needs and considered general dialogue topics.

German LDA. The German LDA model is trained on 20 topics with a corpus of 55,744 documents covering a variety of themes. One topic cluster focuses on messages and questions about the permit of residence for refugees, while another deals with offerings on digital auctioning platforms like eBay Kleinanzeigen. Accommodation is featured in one topic cluster with offerings and expressed needs from locals. Another topic cluster highlights Germans citizens supporting incoming refugees by providing language translation and learning opportunities. Two topic clusters cover the search for accommodation, with a focus on families with children. Another topic cluster revolves around registration questions and searches for help at jobcenters, which are the agencies distributing benefits for refugees. Transportation across borders is the central theme in one topic cluster. Several topic clusters are made up of messages from refugee organizations asking for specific items such as financial donations, clothing, children's clothing, hygiene products, and volunteers to support refugees at common points of arrival, such as train stations. Several topics are general dialogue topics specific to Telegram and are not specific to refugee needs.

BERTopic. Unlike the LDA models, the BERTopic model does not require a specified number of topics before training. Due to memory constraints, we trained the model on a subsample comprised of 557,740 messages, which represents about 30% of the training set with pre-computed embeddings. This model returned 9,162 topics with a minimum size of 10 per cluster, which we reduced the number of topics to 50 after training by iteratively combining the most similar topic representations.

The results of our topic modeling analysis suggest that incoming refugees in Germany have various needs and concerns. These can be broadly clustered into several topics, such as the search for accommodations in Berlin (Topic 2), passports for registration in Germany (Topic 3), access to German language and integration courses (Topic 16), questions regarding children's education and the language barrier (Topic 18), and legal questions about permits to stay, with frequent references to § 24 Aufenthaltsgesetz, the law that allows temporary residence for Ukrainian refugees in Germany (Topic 23). Other topics include travel across Germany and neighboring European countries (Topics 7, 20, 33, and 34), vaccination requirements for COVID-19 and other diseases (Topic 27), and searches for specific support groups for a specific city (Topic 29). Additionally, refugees are concerned with free or discounted transportation through the "9€-Ticket", the risk of scammers and fake housing offers, and healthcare and specialist services (Topics 36 and 42). Clusters of German locals offering housing and help with translation were also identified (Topics 37 and 46). Other topic clusters are generic messages of people thanking each other for helping out with specific questions or specific to the communication medium Telegram, such as channel rules.

Combined Results. Since our method provides a language, a sentiment, and a topic for each message, we can also analyze the predictions together. This enables us to understand better how those variables affect each other. We find that, on average, messages in Russian (-0.026) hold more negative sentiment compared to messages in Ukrainian (0.008). We also find that the sentiment of Russian and Ukrainian messages is most negative in late February and early March, with a more positive and stable sentiment after the middle of March. Here we also find that Ukrainian messages are, on average, more positive than the Russian messages across the timespan of our training set.

We also look at the sentiment for the topics of the BERTopic model. Figure 3 shows the relative sentiment of selected topics that explicitly cover refugee needs. While we find that the majority of messages across all topics are neutral, there is some variation. Topic 19, which covers traveling with pets, pet vaccination, and pet registration, contains a relatively high percentage of negative messages. However, Topics 2 and 25 cover the search for housing in Berlin, and access to free food through the *Tafel*, an organization distributing food to those in need, contain more positive than negative messages.

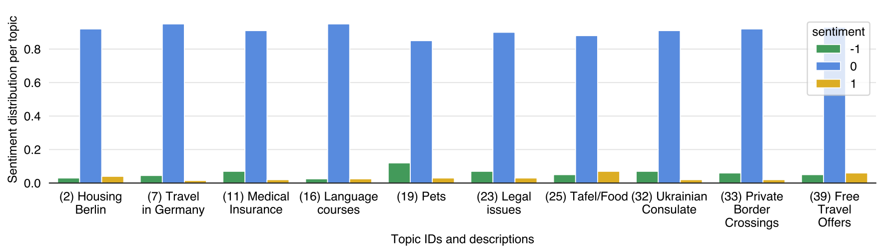


Figure 3. Sentiment distribution for selected refugee needs topics

Dynamic Topic Model. Our dynamic topic model was trained as an extension of the BERTopic model on 16 timestamps, showing the importance of topics over time. Figure 4 tracks the development of Topic 2, which relates to messages on housing needs, primarily in Berlin, with a peak occurrence in early March and a subsequent decline. The peak of the topic frequency in early March, with a decline in the following weeks, indicates that while the need for accommodation, specifically in Berlin, was very high in the early weeks after the start of the invasion, it is not as high anymore.

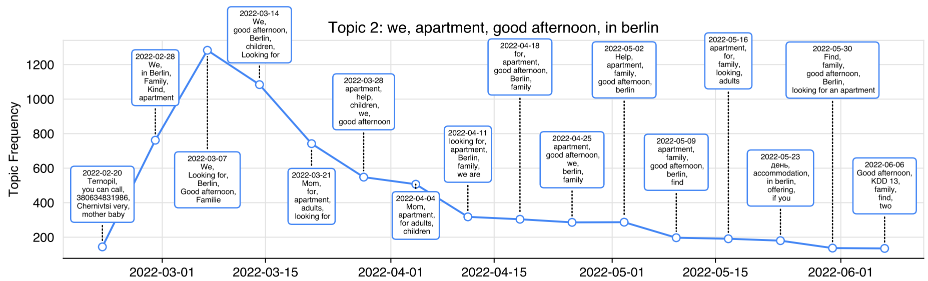


Figure 4. Visualization of topic frequency and representation over time for a selected topic created by Dynamic Topic Model

5.1 Dashboard

We provide an interactive dashboard that can assist stakeholders like policy makers and humanitarian organizations by providing real-time insights into the needs of Ukrainian refugees beyond the timeframe of our training set. As shown in Figure 5, this requires the data collection process to run at intervals, for example, daily. The incoming data are processed, and insights about topic development over time can be generated. We have built and deployed a web-based dashboard that visualizes a subset of that test dataset. The tool's adaptability makes it suitable for various crisis contexts beyond the current one, enhancing its broader application.

6 Conclusion

In this work, we aimed to understand the needs of Ukrainian refugees in Germany by analyzing messages collected from over 130 Telegram channels. We utilized exploratory data analysis techniques along with various Natural Language Processing (NLP) methods, including language identification, sentiment analysis, and topic modeling, to gain insights into the evolving needs of Ukrainian refugees. Our study's methodology and findings contribute to the broader theories of Big Data Analytics and Crisis Communication by demonstrating how social media data can be harnessed to gain rapid insights into populations' needs during crises. This work thereby adds a new dimension to these theories, emphasizing the potential of data analytics in crisis management and communication.

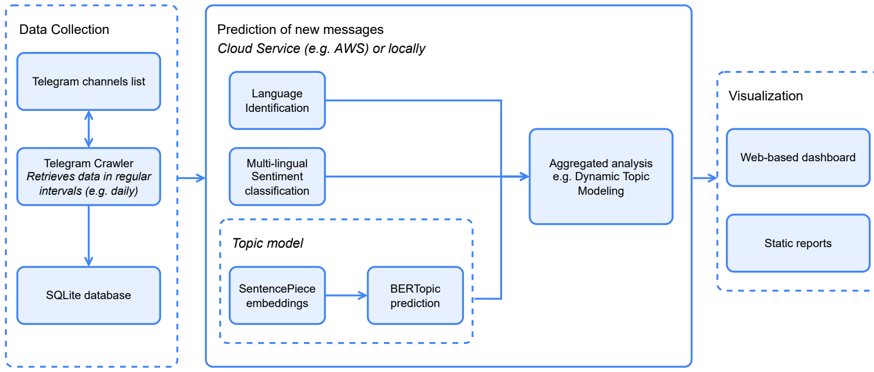


Figure 5. Workflow for real-time analysis

Findings from NLP Analysis. Our NLP analysis identified distinct clusters of refugee needs consistent with expert opinions found in existing literature through both the LDA and BERTopic approaches. These needs included translation assistance, housing assistance, and the need for legal support, registration, and financial assistance. Other needs identified were transportation across borders, language courses, work opportunities, and medical needs. Additionally, our dynamic topic modeling approach enabled us to track the development of these needs over time.

Comparison of Topic Model Approaches. We specifically compared two topic modeling approaches in LDA and BERTopic. Topics generated by the English and German LDA models are messages about needs from the perspective of a humanitarian organization or volunteers, such as offers to assist with translation, calls for donations of goods, or offers for housing. Particularly the Russian LDA model highlighted the need for housing, language courses, legal assistance with jobcenters and registration, and financial assistance. The Ukrainian LDA generated topics on transportation across borders, language courses, and work opportunities. With the BERTopic approach, we could even uncover clusters of housing needs specific to a city or to a group with special needs. We also found clusters that describe the need for information on language and integration courses, along with related topics, job market access, and certificates for education. Topics that cover medical needs are also clearly found with this approach. Also, information needs about legal support for refugee registration, financial welfare, and other administrative questions are found. The need for humanitarian aid items was found mostly through calls from humanitarian organizations to donate those items rather than through refugees themselves.

Refugee Communication via Telegram. Our findings indicate that Ukrainian refugees and local volunteers mainly use Telegram groups to exchange information. This suggests that these channels could improve the German government, authorities, and humanitarian organizations' information distribution. This study highlights the potential for responsible authorities and organizations to address this information gap by interacting with refugees via Telegram groups.

Practical Implications. The practical implications of our research include essential insights into the needs of Ukrainian refugees in Germany during the crisis caused by the Russian invasion. By identifying these needs and tracking their evolution, we have generated a comprehensive understanding of the crisis trends. We also developed a dashboard application to assist domain experts from humanitarian organizations or government entities in accessing insights into the needs of Ukrainian refugees. The information derived from this study can inform strategic decision-making in communication and policy development. The methodology presented in this study can also be applied to similar crises, thereby contributing to evidence-based decision-making during future humanitarian emergencies. Our research techniques also have potential applications in non-crisis contexts, such as understanding social media users' behavior and needs in various domains. These could assist researchers and practitioners beyond this specific domain in designing targeted strategies to meet their constituents' needs.

References

Adema, J. & Maitreyee, G. (2022), 'Following the Online Trail of Ukrainian Refugees through Google Trends', *CESifo Forum* **23**(4).

Akbari, A. & Gabdulhakov, R. (2019), 'Platform Surveillance and Resistance in Iran and Russia : The Case of Telegram', *surveillance and society* **17**, 223–231.

Alyafeai, Z., AlShaibani, M. S. & Ahmad, I. (2020), 'A Survey on Transfer Learning in Natural Language Processing'.

Baldwin, T. & Lui, M. (2010), 'Language Identification: The Long and the Short of the Matter', *Human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics* pp. 229–237.

Barbieri, F., Anke, L. E. & Camacho-Collados, J. (2022), 'XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond', *Proceedings of the LREC, Marseille, France* pp. 20–25.

Basu, M., Shandilya, A., Khosla, P., Ghosh, K. & Ghosh, S. (2019), 'Extracting Resource Needs and Availabilities From Microblogs for Aiding Post-Disaster Relief Operations', *IEEE Transactions on Computational Social Systems* **6**(3), 604–618.

Bergsma, S., McNamee, P., Bagdouri, M., Fink, C. & Wilson, T. (2012), 'Language Identification for Creating Language-Specific Twitter Collections', *Proceedings of the second workshop on language in social media* pp. 65–74.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent Dirichlet Allocation', *Journal of machine Learning research* pp. 993–1022.

Brücker, H. (2022), 'Geflüchtete aus der Ukraine: Eine Einschätzung der Integrationschancen', *IAB-Discussion Paper* .

Brücker, H., Goßner, L., Hauptmann, A., Jaschke, P., Kassam, K., Kosyakova, Y. & Stepanok, I. (2022), 'Die Folgen des Ukraine-Kriegs für Migration und Integration: Eine erste Einschätzung', *IAB-Discussion Paper* .

Bundesministerium des Innern und für die Heimat (2022), Systematische Befragung von 1.936 Geflüchteten aus der Ukraine von 24. bis 29. März 2022, Technical report.

Caprolu, M., Sadighian, A. & Di Pietro, R. (2022), 'Characterizing the 2022 Russo-Ukrainian Conflict Through the Lenses of Aspect-Based Sentiment Analysis: Dataset, Methodology, and Preliminary Findings'.

Chen, E. & Ferrara, E. (2022), 'Tweets in Time of Conflict: A Public Dataset Tracking the Twitter Discourse on the War Between Ukraine and Russia'.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V. (2020), 'Unsupervised Cross-lingual Representation Learning at Scale'.

Dargahi Nobari, A., Reshadatmand, N. & Neshati, M. (2017), Analysis of Telegram, An Instant Messaging Service, *in* 'Proceedings of the 2017 ACM on Conference on Information and Knowledge Management', ACM, Singapore Singapore, pp. 2035–2038.

Debortoli, S., Müller, O., Junglas, I. & vom Brocke, J. (2016), 'Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial', *Communications of the Association for Information Systems* **39**, 110–135.

Egger, R. & Yu, J. (2021), 'Identifying hidden semantic structures in Instagram data: A topic modelling comparison', *Tourism Review* **77**(4), 1234–1246.

Egger, R. & Yu, J. (2022), 'A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts', *Frontiers in Sociology* **7**, 886498.

Gillespie, M., Gillespie, M., Lawrence, A., Cheesman, M., B Faith, Faith, B., Illiou, E., Issa, A., Osseiran, S. & Skleparis, D. (2016), 'Mapping Refugee Media Journeys: Smartphones and Social Media Networks'.

Grootendorst, M. (2022), 'BERTopic: Neural topic modeling with a class-based TF-IDF procedure', *ArXiv* .

Hoseini, M., Melo, P., Júnior, M., Benevenuto, F., Chandrasekaran, B., Feldmann, A. & Zannettou, S. (2020), Demystifying the Messaging Platforms' Ecosystem Through the Lens of Twitter, *in* 'Proceedings of the ACM Internet Measurement Conference', ACM, Virtual Event USA, pp. 345–359.

Imran, M., Castillo, C., Diaz, F. & Vieweg, S. (2015), 'Processing Social Media Messages in Mass Emergency: A Survey', *ACM Computing Surveys* **47**(4), 1–38.

Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2016), 'Bag of Tricks for Efficient Text Classification'.

Jurić, T. (2022), 'Predicting refugee flows from Ukraine with an approach to Big (Crisis) Data: A new opportunity for refugee and humanitarian studies'.

Khatua, A. & Nejdl, W. (2021), Struggle to Settle down! Examining the Voices of Migrants and Refugees on Twitter Platform, *in* 'Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing', CSCW '21, Association for Computing Machinery, New York, NY, USA, pp. 95–98.

Knuutila, A., Herasimenka, A., Bright, J., Nielsen, R. & Howard, P. N. (2020), 'Junk news distribution on Telegram: The visibility of English-language news sources on public Telegram channels', *The Computational Propaganda Research Project* p. 5.

Li, T., Xie, N., Zeng, C., Zhou, W., Zheng, L., Jiang, Y., Yang, Y., Ha, H.-Y., Xue, W., Huang, Y., Chen, S.-C., Navlakha, J. & Iyengar, S. S. (2017), 'Data-Driven Techniques in Disaster Information Management', *ACM Computing Surveys* **50**(1), 1:1–1:45.

Mizrahi, T. & Yallouz, J. (2022), 'Using Internet Measurements to Map the 2022 Ukrainian Refugee Crisis'.

Nazer, T. H., Xue, G., Ji, Y. & Liu, H. (2017), 'Intelligent Disaster Response via Social Media Analysis A Survey', *ACM SIGKDD Explorations Newsletter* **19**(1), 46–59.

Ng, L. H. X. & Loke, J. Y. (2020), 'Analyzing Public Opinion and Misinformation in a COVID-19 Telegram Group Chat', *IEEE Internet Computing* **25**(2), 84–91.

Park, C. Y., Mendelsohn, J., Field, A. & Tsvetkov, Y. (2022), 'VoynaSlov: A Data Set of Russian Social Media Activity during the 2022 Ukraine-Russia War'.

Pohl, J., Seiler, M. V., Assenmacher, D. & Grimme, C. (2022), 'A Twitter Streaming Dataset collected before and after the Onset of the War between Russia and Ukraine in 2022'.

Pötzschke, S., Weiß, B., Hebel, A., Piepenburg, J. G. & Popek, O. (2022), 'Preliminary findings from an online survey of Ukrainian refugees in Germany and Poland'.

Reimers, N. & Gurevych, I. (2019), 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks'.

Sabatovych, I. (2019), ‘Do social media create revolutions? Using Twitter sentiment analysis for predicting the Maidan Revolution in Ukraine’, *Global Media and Communication* **15**(3), 275–283.

Skwarek, J., Nocoń, T. & Kozieł, J. (2022), ‘War in Ukraine - Twitter analysis’.

Smetanin, S. (2020), ‘The Applications of Sentiment Analysis for Russian Language Texts: Current Challenges and Future Perspectives’, *IEEE Access* **8**, 110693–110719.

Trauner, F. & Valodskaite, G. (2022), ‘The EU’s Temporary Protection Regime for Ukrainians: Understanding the Legal and Political Background and Its Implications’, *CESifo Forum* **23**(04), 17–20.

UN Women (2022), ‘Updated displacement survey: Ukrainian refugees and third-country nationals’.

UNHCR (2022*a*), Profiles, Needs & Intentions of Refugees from Ukraine, Technical report.

UNHCR (2022*b*), ‘Situation Ukraine Refugee Situation’, https://data.unhcr.org/en/situations/ukraine.

Urman, A., Ho, J. C. K., Justin Chun-ting Ho & Katz, S. (2021), ‘Analyzing protest mobilization on Telegram: The case of 2019 Anti-Extradition Bill movement in Hong Kong.’, *PLOS ONE* **16**(10).

Urman, A. & Katz, S. (2020), ‘What they do in the shadows: Examining the far-right networks on Telegram’, *Information, Communication & Society* pp. 1–20.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017), ‘Attention Is All You Need’, *Advances in neural information processing systems* **30**.

Walk, E., Garimella, K. & Christia, F. (2023), ‘Displacement and return in the internet Era: Social media for monitoring migration decisions in Northern Syria’, *World Development* **168**, 106268.

Wich, M., Górniak, A., Eder, T., Bartmann, D., CCakici, B. E. & Groh, G. (2021), ‘Introducing an Abusive Language Classification Framework for Telegram to Investigate the German Hater Community’, *ArXiv* .

World Health Organization, & Control, E. C. f. D. P. (2022), Early warning, alert and response to acute public health events in refugee-hosting countries: Ukraine crisis, Technical Report WHO/EURO:2022-5506-45271-64773, World Health Organization. Regional Office for Europe.

Yigitcanlar, T., Kankanamge, N., Preston, A., Gill, P. S., Rezayee, M., Ostadnia, M., Xia, B. & Ioppolo, G. (2020), ‘How can social media analytics assist authorities in pandemic-related policy decisions? Insights from Australian states and territories’, *Health Information Science and Systems* **8**(1), 37.

Zhu, Y., Haq, E.-u., Lee, L.-H., Tyson, G. & Hui, P. (2022), ‘A Reddit Dataset for the Russo-Ukrainian Conflict in 2022’.