Data For Ukraine

Project Information Summary

Updated: Thursday, March 24, 2022

In critical moments like the Russian invasion of Ukraine, as scholars it is our duty to use our technical skills, area knowledge, and team capacity to produce research that supports citizens and policy makers. Data For Ukraine is our collective effort to do just this.

Data For Ukraine (#DataForUkraine) is a major international collaboration between scholars at The Kyiv School Of Economics, the MOBILISE project at the Universities of Manchester and Oxford (led by Olga Onuch), the Machine Learning for Peace project at Duke University (led by Erik Wibbles), political scientists at the University of North Carolina (led by Graeme Robertson) and the Inter-Disciplinary Lab for Computational Social Science at University of Maryland (led by Ernesto Calvo).

The aim of the project is to gather data on civilian resistance (CR), human rights abuses (HRA), displacement of people (IDP) and humanitarian support/needs (HS) across Ukraine, with the intention of providing timely information for private citizens, NGOs, INGOs and policymakers responding to the Russian invasion and war. To do so, we use hourly data from the Twitter API to report on the incidence of important events.

The data are drawn from an initial collection of more than 400 Twitter accounts (and their followers) covering politicians, civil society activists, journalists and media at the national and local level all across Ukraine and including as broad a range of political positions as possible. Using this initial batch, we identify key networks/communities of accounts and identify key nodes of information. The list of accounts is regularly reviewed to maximize relevant content.

Tweets are automatically searched for more than 600 keywords in Ukrainian and Russian. Keywords were initially derived iteratively from social science theory and translated into Ukrainian and Russian by native speakers. Through a period of consistent multi-team member verification and careful study of the living language actually used in Tweets about the phenomena of concern, we expanded the terms to include these ‘real world’ formulations. Once collated and collected, the keywords are classified into 4 categories with the goal of separately identifying civilian resistance (CR), human rights abuses (HRA), displacement of people (IDP) and humanitarian support/needs (HS). Following the collection of the tweet data, repeated human manual verification of a random sample of Tweets is used to ensure and improve the accuracy of classification of events. Our team is committed to continue and improve the verification through the data collection process as and when possible.

The results are displayed on the Data for Ukraine website (https://mlp.trinity.duke.edu/dataforukraine.php#en).

Currently the website presents data on the change in rate of reports amounting to the occurrence of major event (the spikes) over the previous three days. The data, and the graphics depicting it, is updated every three hours. Data are presented for each of the four categories of events and are also classified by rayon and presented on a map. This allows users to identify what the changes in the rates are, which events are being reported, and to track where events are occurring.

In our initial assessment, we find our machine learning approach helps identify major events of human rights abuses roughly 3hrs before media are able to report on them. For each major
spike we can identify the level of change in rates of reports about a human rights abuse, the magnitude of the event compared to all events of its kind, and also where in Ukraine the event is occurring.

**As an example**, in our tracking of Civilian Resistance and Human Rights Abuses (graph for CR shown) on the 21 March we were able to immediately identify the beginning of a major event (the spike) just as the very first reports came in. This major event was the Russian occupational forces shooting at peaceful protesters in Kherson (in addition to reliance on our modelling we also verified the twitter text and image data manually).

- This fairly immediate capturing of a major incident of Civilian Resistance and Human Rights Abuses in real-time as it unfolded is particularly important for monitoring purposes.
- But the detailed data we collect can also be revisited for more forensic analysis of such major events post fact allowing policy makers, journalists, and activists to make reports on Civilian Resistance and Human Rights Abuses but also track and trace these major events for future court filings.

Work on the public-facing website is ongoing as we refine presentation of our data. We will also shortly begin providing **weekly analytical reports** covering and assessing any patterns in event occurrence and their significance. We will employ our projects’/labs’ previous sources of data on key geospatially varied factors in Ukraine in order to triangulate and better explain key changes in the occurrence of major CR, HRA, IDP, HS events on the ground that they reflect or correlate to. Our team can also provide specialized/focused reports for policy makers.

Our team are keen on working Human Rights experts, journalists, and policy makers to share this information in real-time whilst also keeping a detailed record for future use in courts, proceedings, and research.

Please feel free to contact us for further information:

Olga Onuch [olga.onuch@manchester.ac.uk](mailto:olga.onuch@manchester.ac.uk)
Graeme Robertson [graeme@email.unc.edu](mailto:graeme@email.unc.edu)
Erik Wibbels [e.wibbels@duke.edu](mailto:e.wibbels@duke.edu)
Ernesto F. Calvo [ecalvo@umd.edu](mailto:ecalvo@umd.edu)