Oxford Policy Management

Methodological Note and Tools

Inclusive Education Evaluation by using Big Data Sources of Information

Oxford Policy Management

August 2023

About Oxford Policy Management

Our vision is for fair public policy that benefits both people and the planet. Our purpose is to improve lives through sustainable policy change in low- and middle-income countries.

Through our global network of offices, we work in partnership with national stakeholders and decision makers to research, design, implement and evaluate impactful public policy. We work in all areas of economic and social policy and governance, including health, finance, education, climate change and public sector management. We have cross-cutting expertise in our dedicated teams of monitoring and evaluation, political economy analysis, statistics, and research methods specialists. We draw on our local and international sector experts to provide the very best evidence-based support.

Oxford Policy Management Limited
Registered in England: 3122495

Level 3, Clarendon House
52 Cornmarket Street
Oxford, OX1 3HJ
United Kingdom

Tel: +44 (0) 1865 207 300
Fax: +44 (0) 1865 207 301
Email: admin@opml.co.uk
Website: www.opml.co.uk
Twitter: @OPMglobal
Facebook: @OPMglobal
YouTube: @OPMglobal
LinkedIn: @OPMglobal

Preface

This study was carried out by Oxford Policy Management (OPM) between January and July 2023. The project manager was Udayan Rathore and the remaining team members included Ida Brzezinska (Data Analyst), Arseniy Gurin (Senior Data Scientist), Paul Jasper (Team Leader), Elayn Sammon (Senior Inclusion Policy Expert), Zhanar Zhaxylykova (National Evaluation Expert). For further information contact Udayan Rathore at Udayan.rathore@opml.co.uk.

The contact point for the client is Udayan Rathore and can be reached out to at Udayan.rathore@opml.co.uk. The client reference number for the project is A-5421.

The implementation of this study 'Inclusive Education Evaluation by using Big Data Sources' would not have been possible without extensive contributions from a range of organisations and people. We are grateful to the officials / experts of UNICEF Kazakhstan for trusting us with this study and providing the financial support to undertake this research. We would like to thank Tatiana Aderkhina and Raushan Ibrasheva for their inputs at the inception phase of the project as well as their comments on the Preliminary Findings presentation. We are also thankful to the Junction Bulgaria team whose work on the formative evaluation provided us with valuable insights on the inclusive education landscape in Kazakhstan. Also, we would like to express our gratitude to the reviewers of the inception phase presentation at UNICEF.

We are grateful to all the individuals and organisations who posted their perspectives on inclusive education in Kazakhstan on publicly accessible social media platforms, without which this research would not have been possible.

We would also like to acknowledge the contributions of OPM staff Varun Vashistha, and Vladimir Imnaishvili in managing the administrative and financial aspects of this study. We would also like to thank Matilda Eriksson for assisting us with operational details of Meltwater, which was used for analysing one component of social media data in the study.

All opinions expressed, and any mistakes in the report remain the responsibility of the OPM team.

Oxford Policy Management Limited
Registered in England: 3122495

Level 3, Clarendon House
52 Cornmarket Street
Oxford, OX1 3HJ
United Kingdom

Tel: +44 (0) 1865 207 300
Fax: +44 (0) 1865 207 301
Email: admin@opml.co.uk
Website: www.opml.co.uk
Twitter: @OPMglobal
Facebook: @OPMglobal
YouTube: @OPMglobal
LinkedIn: @OPMglobal

Table of contents

List of tables, figures, and boxes

Introduction

This document forms part of the Final Package submission and describes the digital setup and methodological approach to the analysis undertaken in the "Inclusive Education Evaluation by using Big Data Sources of Information" study. The broader context and overall methodological approach to this study are presented in the 'Descriptive Analysis – Final Report' document that was submitted as part of this project.

The present document consists of three parts:

1) A methodological note, which contains a technical description of our approaches to social media listening (SML) and analytical methods used with the Meltwater platform as well as self-programmed scrapers.[1] This note is mainly based on our description of study methods presented in the beforementioned 'Final Report'.

2) Data architecture, containing a verbal description of the Python scripts, folder structure, and expected outputs used with the self-programmed scrapers method – this is accompanied by a submission of the code and raw data in separate files. In the case of Meltwater, since raw data cannot be downloaded from the platform, we provide a description of how our search queries were used for analysis and data visualisation.

3) Finally, the third part contains recommendations on making our analysis sustainable. This section discusses lessons learnt from our experience using SML tools and puts forward a set of recommendations for adopting the methodology in the context of similar analyses in the future.

As mentioned before, overall this report is based primarily on the sections on methodology from our "Inclusive Education Evaluation by using Big Data Sources of Information" final report, which provides a detailed description of methods and the use of SML tools. Given this, some sections will have little or no changes to that report, except for additional detail in descriptions of methodology.

[1] Meltwater is a social media listening platform (https://www.meltwater.com/en). Last accessed on July 10, 2023.

2. A methodological note

2.1. Scraping and processing text data from the Internet

In order to answer the evaluation questions (EQs) of interest, we relied on social media listening tools. We applied two approaches, depending on the stakeholder group whose online discussions we wanted to capture, resulting in two text corpora, i.e. two different collections of text data:

i. **Corpus 1** was created using Meltwater, an online platform which offers scraping and analytics of social media and online news data. This collection of text data is intended to capture the **discussions of the general online public** on inclusive education in Kazakhstan. We explain how we scrape – i.e. 'collect' – text data using Meltwater in section 2.1.1. df

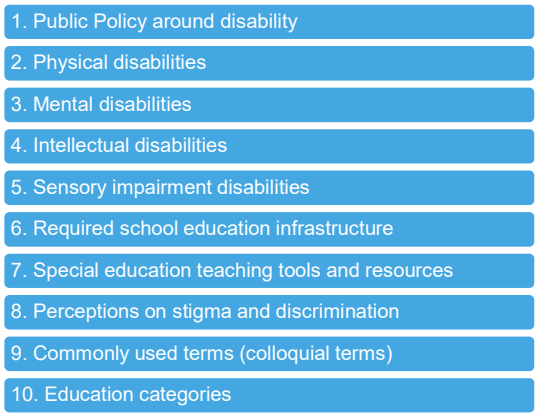
ii. **Corpus 2** was built using self-programmed scrapers and pulled only media **discussions of specific organisations** who are active in the inclusive education space in Kazakhstan. We explain how we gain access to this data in section 2.1.2.

Through these complementary approaches of analysing similar EQs from different perspectives, we triangulated insights as well as contrasted these findings across different stakeholders. Importantly, to study barriers that impede inclusive education, we again used both these on a different set of data that is scraped using a modified search query to incorporate barriers to inclusive education (see Figure 1).

2.1.1. Corpus 1: Using Meltwater

As a first step in building a search query in Meltwater, we identified multiple sub-topics within inclusive education through a literature review and expert discussions, shown in Figure 1 below. For each sub-topic, we created a combination of keywords in Russian and Kazakh that accurately described the chosen topic and considered the specifics of each language and different stakeholder groups. When compiling our keywords, we took into account declination and conjugation in Russian and Kazakh in order to encompass all possible variations of words frequently used in media dialogues. An example of key words for Sub-topic 2 "Physical disabilities" is shown in Box 1 below.

Figure 1: Sub-topics within Inclusive Education



Box 1: Key words for Sub-topic 2: Physical disabilities

Wheelchair user, muscular dystrophy, spina bifida, cerebral palsy, diseases of the musculoskeletal system, lameness, dwarfism, epilepsy, poor physical coordination, chronic pain, physical handicap, diseases of the central or peripheral neurological systems.

Once our sub-topics and keywords were defined, we initiated search queries within the Meltwater platform, scraping online discussions on inclusive education in Kazakhstan. We used the "combined search" option, which allows us to combine multiple categories of search queries in Meltwater using Boolean operators, i.e. words and symbols such as "AND", "OR" that expand or narrow the search query according to specific parameters. For instance, if we were interested in running a search that includes both categories 1 and 2 from Figure 1**, we would combine two sets of keywords using the "AND" operator. The specific keywords as well as the use of the Boolean operators in Meltwater are illustrated in** Figure 6 **for Query 1 on Inclusive Education KZ and in** Figure 7 **for Query 2 on Inclusive Education RU in the Annex.** Meltwater covers the following sources: news, blogs, Facebook, Instagram, Twitter, Forums, LinkedIn, Pinterest, Reddit, RSS, Tik Tok, Twitch, YouTube. Importantly, we note in the context of Instagram, Meltwater does not allow one to extract the whole text of each post. Thus, this is exercise is undertaken in Corpus 2, discussed in the subsequent section (2.1.2).

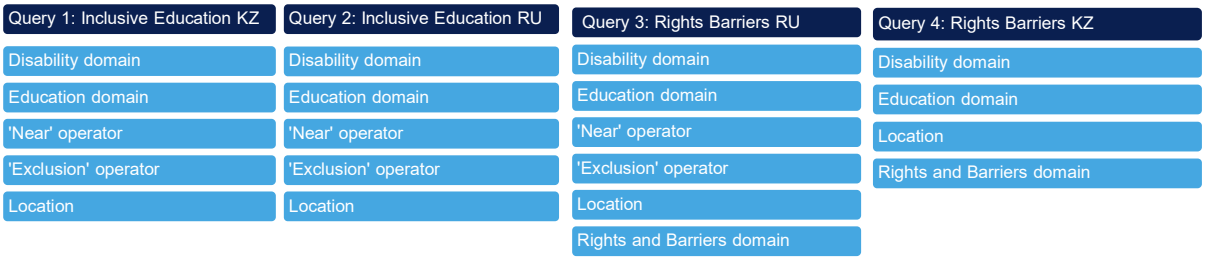
To test how accurate these search results are and improve them wherever necessary, we undertook a validation process, where we initially ran our search queries for a period of 90 days only and assessed the relevance of our results to the topic. To ensure accuracy, we randomly selected a portion of the results and thoroughly read each post or news item, classifying them as either "relevant" or "not relevant" to the topic of 'inclusive education'. Multiple iterations of the search queries were conducted to achieve a high percentage of relevant outcomes. Through this process, we were able to increase the number of relevant results from 30% to 77% in the case of searches in Russian, and from 6% to 80% in the case of searches in Kazakh. Once a high level of relevance was attained, the established search queries were utilised to collect data over a one-year period, specifically from **May 1, 2022, to May 1, 2023**.

Our final search queries can be divided into two categories, both of which were ran in Russian as well as Kazakh. The breakdown of specific sub-topics contained within these categories is shown in Figure 2.

1. **Inclusive Education:** combined queries related to sub-topics of "Disability" and "Education". These queries incorporated relevant keywords that specifically covered topics 2, 3, 4, 5, 6, and 10 from the list of sub-topics in Figure 1 above. Additionally, in order to limit our analysis to results in Kazakhstan, we included a "Location" query that specified our results must mention locations within Kazakhstan. To enhance the quality and relevance of our data, we implemented the "Exclude" function specifically designed to filter out spam websites. Finally, the "NEAR" operator ensures that all relevant keywords are within a specific proximity to each other, for instance no more than 50 words apart.

2. **Rights and Barriers:** contains the sub-topics of "Disability" and "Education within Kazakhstan, but additionally includes queries specifically focused on the topic 8 "Perceptions on stigma and discrimination" from Figure 1, defined as "Rights and Barriers domain" search in Figure 2. This category covers keywords related to stigma and discrimination, as well as violence and violation of rights - in order to gain insights into potential challenges and barriers faced by children with spatial educational needs. The slight differences in how our "Rights and Barriers" query is defined in Russian and Kazakh emerged from the validation process, as in each iteration we selected the query that gave the highest proportion of relevant results. Importantly, as this category is built within the topic of "Disability and Education", some of the themes that feature prominently as topics of discussion may, by construction, refer to barriers in inclusive education.

Figure 2: Key sub-topics in our final search queries



Furthermore, we incorporated queries specifically focused on gender, which were not included in the previous topics. These queries consisted of separate searches for gendered keywords associated with girls and boys within Meltwater. However, we discovered a low relevance of the collected publications regarding gender issues. Unfortunately, the publications did not contain specific information or substantial insights related to gender in the context of inclusive education in Kazakhstan.

The number of results in our final search queries for all four categories from Figure 2 are shown in Table 1 below. Query 2, ran in Russian, yielded the largest number of mentions of the topic of inclusive education, with a total of 1,200 mentions. Query 1, ran in Kazakh, was mentioned within online discussions 515 times. Given Rights Barriers is a more specific search, intended to pick up discussions focused on stigma, discrimination, and violation of rights of children with special education needs – we see a lower number of mentions than

Inclusive Education. As shown by Queries 3 and 4, the topic of Rights Barriers has more mentions in Russian than in Kazakh, with 326 and 258 total mentions, respectively.

Table 1: Number of results from key search queries

Search query	Query 1: Inclusive Education KZ	Query 2: Inclusive Education RU	Query 3: Rights Barriers RU	Query 4: Rights Barriers KZ
Number of Mentions	515	1,200	326	258

We provide more details on how our search queries are displayed and stored in Meltwater in Section 3.1, including the issue of the raw data architecture, which cannot be downloaded from Meltwater. In addition, we describe the use of analytical tools and data visualisations through dashboards available in the platform.

2.1.2. Corpus 2: Using self-programmed scrapers

In addition to capturing perspectives of general public, we also looked at specific organisation working on inclusive education in Kazakhstan. By virtue of their experience and institutional knowledge on inclusive education, their perspectives may differ from those of the general audience and are thus an integral part of the study. The full list of organisations whose online discussions we scraped can be found in Table 3 in the Annex.

Through a purposive research design, we identified multiple organisations and their publicly accessible social media handles and websites, which were scraped via a self-written programme to create text Corpus 2. This means that we wrote a computer programme using the programming language 'Python', which accessed the social media sites and websites and downloaded publicly accessible text. A major component of text Corpus 2 came from Instagram channels, Telegram groups and channels, and web sources such as blogs and news outlets. Importantly, this 'manual' scraping approach was also beneficial as the Meltwater approach presented above did not scrape most of the above sources except Instagram. Moreover, manual scraping of Instagram channels also allowed us to extract the whole text of each post thereby facilitating a more in-depth analysis.

To account for exclusivity of discussions, we classified each source either as dedicated or not. Dedicated sources undertook discussions exclusively on disability and education whereas the second category shared perspectives on other topics as well. As manual scraping allowed for data extraction over longer time durations, messages posted between **January 1, 2022** to **April 30, 2023** were included in the analysis, which means that this timeframe is slightly different than the one for Corpus 1. Moreover, we only selected and scraped channels and sources that were identified as part of the purposive design. We applied the following scraping processes to each of the data sources for Corpus 2.

Instagram: Here, we used an external tool called Apify (link) to perform data scraping. Instagram is a closed platform, requiring specialised tools to undertake data extraction, which is often a difficult task. This data was provided in JSON format. From this, we extracted selected fields such as the post text itself, the location of the post (when available), the message, author. We anonymised any identification information after analysis.

Telegram: We downloaded this data using the telegram desktop application. The application does not distinguish between a channel or a group and therefore we applied the same scraping method. We saved the data to the disk in an HTML format, which was then processed to extract each individual message and the message date.

Web sources: Since online sources vary considerably in their structure, we wrote a separate scraper for each of the web sources. For each source, we first extracted all articles, i.e. all text published on the website. As a second step, we extracted the actual article content together with the date.

It should be noted, again, that all data scraped using the above processes was publicly available text data. This means that we did not scrape any private messages or text published in private groups. Once we completed the data scraping, we saved the results were in a local file and further processed it. This involved filtering the entire text corpus by date and translating all the text content to Russian.[2] We used the Yandex translator API for this task (link).

To keep the analysis consistent across the two text corpora, we applied the same queries as in the Meltwater search to each record. In particular, this yielded the following filter: a record is considered relevant for the analysis if it comes from a dedicated source or has mentions of both disability and education. From a total of ~12,000 records, ~10,500 records were deemed relevant by this filter. Also, we identified about 2300 and 3400 key searches on Disability and Education, respectively. Importantly, following these steps, we identified 750 search queries on barriers to inclusive education (Table 2). Here, we note that while for Corpus 1 search queries determined the resulting records themselves, in Corpus 2 they were used to select a subset of the already scraped documents. Since analysis on Corpus 2 was performed in Russian language only, there is no need to specify the language: if the original document was not in Russian, we translated it first and then applied the query.

Next, we applied the NLP model to remove stop words and lemmatize words in the text. This was necessary in order to perform the next steps of data processing. Wherever available, geographical locations mentions were extracted from the text.

Table 2: Number of results from key search queries

Search query	Total records scraped from 1st Jan 2022 to 30 April 2023	Dedicated source	Disability	Education	Barriers
Number of Mentions	11891	10181	2277	3424	753

[2] Due to the purposive selection of sources, the text data was in commonly used languages in the region, which are either Kazakh or Russian. The decision to translate to Russian was taken because further processing steps required a trained Natural Language Processing (NLP) model to exist for the language. Unfortunately, the library that we were most familiar with in this context (Spacy (link)) did not include a model for the Kazakh language, and a model could not be found in other common libraries.

2.2. Data analysis

2.2.1. Corpus 1

Given text data from Corpus 1 was collected using the Meltwater platform, we relied primarily on analytical tools available within Meltwater to conduct data analysis. For both our general search query on Inclusive Education as well as the specific query analysing the rights and barriers of children with special educational needs (see Figure 2), we used the following methods:

1. **Mentions trend:** This shows how the volume of posts mentioning our topics of interest evolved over time (for instance, monthly) during the study period. This method allows us to identify spikes of mentions at a time when online discussion on inclusive education in Kazakhstan was especially active, or conversely – quiet, and associate those with particular events.

2. **Top publications by reach**: Reach estimates the potential viewership of any particular article based on the number of monthly unique visitors to the specific source. This feature allowed us to analyse which publishing sites had the largest number of viewers, giving us insight into where most widely viewed discussions on inclusive educations are hosted, as well as the ownership of this source.

3. **Top news:** This allowed us to see the content of posts that had the largest reach, and thus identify the most widely read online publications relating to inclusive education.

4. **Top sources:** This feature displays the most popular sources for our search queries and included all platforms covered by Meltwater, i.e. news, blogs, Facebook, Instagram, Twitter, Forums, LinkedIn, Pinterest, Reddit, RSS, Tik Tok, Twitch, YouTube.

5. **Sentiment analysis:** Using Meltwater's Natural Language Processing algorithm, this method allowed us to determine whether publications have a negative, positive, or neutral tone. We also combined this feature with method 3 "Top news" to display top news that were assigned a particular sentiment.

6. **Hashtag analysis:** This method allowed us to view the most commonly used hashtags as well as the number of posts that have been tagged to each hashtag.

7. **Key words**: This allowed us to view the most commonly occurring words, together with their frequency. We manually excluded "stop words", such as "and" or "the" that did not add any meaningful context to our results.

8. **Gender analysis**: To check if one group (girls or boys) got a disproportionately higher representation over the other in discussions on inclusive education (disability and education), we compared the shares of messages from each group across the text repositories of "education and disability" and "education", respectively. The latter category did not include "disability" and thus served as a reference category. To check for any differences in sentiments across boys and girls, we also conducted a sentiment analysis by gender for both the text corpora.

All of the above-mentioned analytical tools are applied within the Meltwater platform in a 'click and select' way. This means there is no associated code for this analysis. Similarly, all social media data resulting from search queries is stored within Meltwater, which does not allow for downloads of the raw data corpus. Any future replications of this analysis will therefore need to be carried out within the SML platform – either Meltwater or another platform, for instance TalkWalker.

2.2.2. Corpus 2

We used multiple analytical methods to study the text in Corpus 2 extracted from specific organisations working in the inclusive education space in Kazakhstan. These, alongside their purpose are listed below.

1. **Number of messages per date as trend:** Similar to mentions trend above, these facilitated studying trends of discussion on topics of inclusive education and how they varied with the academic session.

2. **Word frequency analysis:** This analysis provided a visual snapshot of 50 most common meaningful words that featured in the discussions on inclusive education.

3. **Topic modelling using Latent Dirichlet Allocation:** Topic modelling is a statistical exercise that allows us to identify cluster of similar words within a body of text. This analysis cannot be executed on Meltwater and was exclusively undertaken for Corpus 2. This allows to identify the underlying topics in the corpora of documents. It provides a way to represent each document as a distribution over the identified topics.

4. **Topic evolution over time:** An extension of the approach above, this method reflects how discussions under each of the thematic area above varied over time.

5. **Messages per country or region normalized by population:** In cases where geo-locations could be identified, this analysis provided a geo-spatial view of the engagement on these discussions.

6. **State programme mentions:** To check if certain programmes were being discussed more than others on social media, we prepared a comprehensive list of active programmes and checked if they find any mention in the text corpus.

7. **Gender analysis:** As done for Corpus-1 above, we also undertake a gender based analysis in Corpus 2.

Importantly, to study barriers that impede inclusive education, we again used the methods stated above but on a different set of data that is scraped using a modified search query to incorporate barriers to inclusive education (see Figure 1).

The bulk of the analysis described above is carried out in the *'to_silver.py'* Python script (see Section 3.2 on details on the data architecture). Scraped data is filtered, translated to Russian, enriched with Meltwater queries results, location is extracted, NLP applied and topic modelling is applied to the whole dataset as well as selected subsets. In terms of individual analytical methods, the *'topic_analysis.py'* file contains functions used to perform topic modelling and the *'nlp.py'* file contains functions that are responsible for running Natural Language Processing (NLP) tasks. The *'readme.txt'* file in the ZIP file *'unicef_kaz'* that accompanies this submission describes the code structure in more detail.

2.3. Methodological limitations

We applied innovative social media listening tools as our main methodology for undertaking this evaluation. While the use of big data offers an opportunity to include online discussions that would not feature within traditional survey-based evaluations, this approach carries its own methodological limitations. In particular, we note the following:

1. **Instagram and Facebook search queries in Meltwater**: There are limitations to the breadth of search queries for Instagram and Facebook in Meltwater, possibly due to data privacy policies of these social media platforms. In the context of Instagram, we were only able to view posts of business accounts that we specifically tracked. These were added manually by selecting up to 30 hashtags and up to 30 accounts. For Facebook, we are able to view all public discussions, including comments. However, if we wanted to track the activity of specific pages, there is a limit of 100 accounts. Therefore, it is possible that only a very small subset of the discussions on these platforms was picked up by Meltwater. We address this limitation in Corpus 2, where we used self-programmed scrapers to access Instagram posts directly.

2. **Meltwater analytical tools**: While the online platform offers a wide range of analytical capabilities, outlined in section 2.2.1, it did not allow the user to download the raw text corpus from its online searches or view the post in full, thus limiting our ability to conduct independent analysis. As a result, the analysis undertaken on Corpus 1 had to be restricted to the tools that Meltwater offers, with no ability to validate the results using our own methods. This was particularly limiting in the context of gender analysis for Corpus 1, where we were restricted to adding key words associated with boys and girls to our search, as opposed to methods such as dictionary search, which could have been possible had we had access to the raw data corpus.

3. **Geo-location**: Given the location of most posts is unknown or not disclosed by the user due to a variety of considerations, including privacy, we relied on key words that mention locations within Kazakhstan to ensure that the online discussion concerns inclusive education *in* Kazakhstan specifically. For the same reason, we were limited in our ability to carry out an analysis of the spatial distribution of online conversations on inclusive education within Kazakhstan. Despite this, we provide a map of locations of all posts that have been geo-tagged for Corpus 2.

4. **Relevance of results**: We relied on a validation process during which a human expert draws a random sub-sample of results from our search queries, and determines what percentage can be considered relevant to the discussion on inclusive education in Kazakhstan. Given that reviewing several hundreds of posts individually would be infeasible, we drew conclusions on relevance on the basis of this sub-sample, rather than the full Corpus. An alternative approach could have been to train a machine learning algorithm to identify the concept of “inclusive education in Kazakhstan” in text, and classify *all* posts as either relevant or non-relevant – however, this method comes with its own set of considerations, such as bias and accuracy in the algorithm, as well as a large training sample required.

5. **Language translations**: For Corpus 2, we used a pre-trained Natural Language Processing model to clean the raw text that was scraped from the web, as well as to perform topic modelling. Given that Kazakh language was unavailable in the Python library that we were most familiar with (SpaCy), and could not easily be found among

other libraries, we relied on the Yandex translator API (link) to translate all Kazakh content into Russian before analysis. As a result, slight differences in meaning, particularly concerning vocabulary that has a specific cultural connotation, might occur.

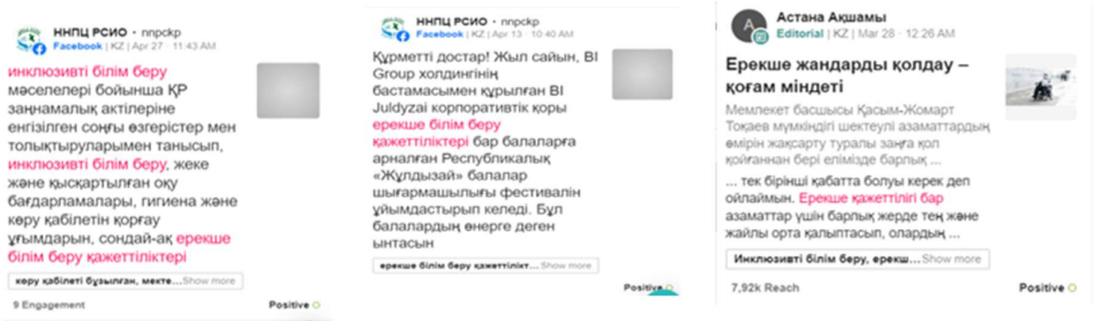
3. Data architecture

As mentioned above, we adopt two approaches to social media listening in our analysis in the 'Inclusive Education Evaluation using Big Data Sources of Information' project. As discussed in the report, we execute finalised search queries to obtain two text corpora. We rely on Meltwater to create Corpus 1, which represents discussions of the general online public. In order to build Corpus 2, which consists of text data for specific organisations who were active in the inclusive education space in Kazakhstan, we used self-programmed scrapers in Python.[3] The use of Python scripts allowed us to download the raw text data, which we share separately as a compressed ZIP format file titled *'unicef_kaz.zip'.* On the other hand, Meltwater provides an integrated environment to search, analyse, and visualise data, but does not allow for raw data downloads. Thus, for Meltwater, we only share details on the construction of search queries and how these were used for data analysis and visualisation.

3.1. Data architecture using Meltwater

Meltwater is a social media listening platform that offers a wide range of functionalities, including constructing search queries, scraping, analytical tools described in Section 2.2.1, data visualisations, and dashboards. It does not, however, allow for downloads of raw text data for independent analysis. For that reason, we are unable to provide the raw data architecture from Meltwater. An example of how social media posts are displayed in Meltwater is shown in Figure 3 below, which shows top positive news with the biggest reach for our Query 1 on Inclusive Education (KZ). Posts are displayed with key words highlighted in pink. This view often displays only the text that is in close proximity to the keyword and the user needs to follow an external link to a particular social media site in order to view the post in full. Although it is possible to download a csv file in Meltwater, which contains information about the post such as its authors, source, and key words – the full content of the post cannot be downloaded, making this format unsuitable for independent analysis.

Figure 3: Top positive news with the biggest reach

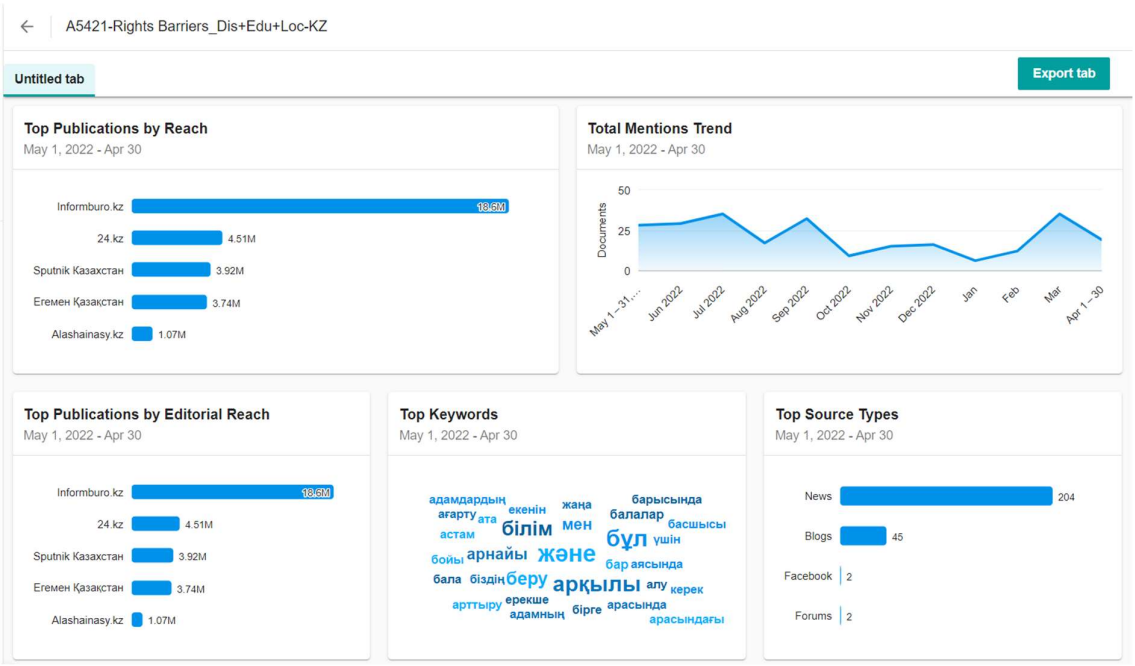


Given the wide range of analytical tools available in Meltwater, we carry out most of our analysis for Corpus 1 within the platform. The first step is a construction of search queries based on our key words related to topics of inclusive education in Kazakhstan. We describe

[3] Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation (https://www.python.org/about/). Last accessed August 12, 2023.

how those are created in detail in Section 2.1.1. Results from search queries can be saved within Meltwater and organised using tags. This data is then stored in the platform and can be used for analysis. In terms of data visualisations, Meltwater allows the user to easily display visual results from analysis on a dashboard. An example of a dashboard generated in Meltwater for our Query 4 on Rights and Barriers KZ is shown in Figure 4 below. It automatically generates data visualisations resulting from our key analytical methods: top publications by reach, mentions trend, top keywords, and types of social media sources for our search query.

Figure 4: Meltwater dashboard for Query 4: Rights Barriers KZ



3.2. Data architecture using Python

The ZIP format file titled *'unicef_kaz'* contains the code that was used to scrape a selection of sources and to further process the resulting data. Within the *'unicef_kaz'* file, there is a readme.txt file that provides further information on each Python script, data structure, and expected outcomes. The resulting data folder contains the scraped data divided in three levels:

1. **Bronze:** bronze data is unprocessed data or data that was only processed in order to be saved in tabular format.

2. **Silver:** silver data is data that has been cleaned, and had other transformations applied to enrich it.

3. **Gold:** data in this folder is the finally aggregated data ready for user consumption and analysis (charts).

Three main data sources are used:

1. **Instagram:** Instagram is a social media sharing and social networking application that allows its users to upload media in various formats. This can be edited and organised by thematic tags (hashtags).

2. **Telegram:** Telegram Messenger is a globally accessible freemium, encrypted, cloud-based and centralized instant messaging (IM) service.

3. **Web:** We also used webpages of specific organisations working in the inclusive education landscape in Kazakhstan.

Due to the different nature of the sources, initial scraping is done differently for each of these three sources. We scraped selected Instagram channels using a paid resource (apify.com), which gave the contents in the JSON format. In order to scrape Telegram channels and groups, we used the telegram desktop application, which allows to download the whole history for the selected source. In terms of web sources – we scraped those directly. The code used to scrape and to perform a preprocessing of the scraped data is contained in the respective folders whose layout is described in Figure 5 below.

Figure 5: Layout of data files

Scraped raw data is saved to the folder at the following path: *'/data/bronze/[source]/raw_data'*. Similarly, scraped data after preprocessing is saved to the folder: *'/data/bronze/[source]/processed_data'*. After the scraping and preprocessing is performed, data can be processed using the scripts in the respective folder titled *"/processing*".

This last folder contains a variety of files. A description for each is given below:

- **common.py:** contains helper functions often referenced by multiple files, such as data save and load functions and others.

- **const.py:** contains constant variables definitions.

- **extract_location.py**: contains functions and logic responsible for extracting the location from the scraped data.

- **meltwater.py:** contains the code that allows to run queries as defined in Meltwater on the scraped data. Meltwater queries are stored in the subfolder: *'/processing/meltwater_queries'*.

- **nlp.py:** contains functions that are responsible for running Natural Language Processing (NLP) tasks on scraped data

- **to_bronze.py:** contains functions that gather all scraped data and save it to a single file in tabular format in the 'bronze' data folder.

- **to_gold.py:** functions in this folder produce charts that display aggregated data for user consumption.

- **to_silver.py:** this is the core of the processing pipeline. Bronze data is filtered, translated to Russian, enriched with Meltwater queries results, location is extracted, NLP applied and topic modelling is applied to the whole dataset as well as selected subsets.

- **topic_analysis.py:** this file defines functions used to perform topic modelling on the scraped data.

- **translation.py:** this file contains functions used for translating the scraped data to Russian. The choice of language depends on the fact that no free/readily available NLP pipelines were found at the time of writing the scripts for Kazakh language. This file relies on private API keys that are not included in this distribution.

In summary, first the scraping scripts are applied, if no data is present.

As mentioned, only web sources are scraped using python scripts, instagram and telegram are scraped manually and no scraping scripts exits for those.

To run web scraping, one needs to execute all the scripts in the folder /web/scraping.

After the raw data is saved in folder defined in this file, one needs to execute the data preprocessing scripts:

/instagram/processing/main.py

/telegram/processing/main.py

/web/processing/<all except common.py>

Finally, the following scripts from folder /processing need to be executed in the following order. We note that these three scripts call each other sequentially and thus there is no need to execute them individually.

1) to_bronze.py

2) to_silver.py

3) to_gold.py

4. Recommendations on making the tool sustainable

The methods described in this note present two sets of social media listening (SML) tools: 1) Meltwater - an online platform which offers scraping and analytics of social media and online news data; and 2) self-programmed scrapers that pull media discussion from specific organisations and platforms. Those tools were used in our evaluation to inform the state of online discussion on inclusive education in Kazakhstan, but can serve as a general reference to guide similar types of descriptive analyses in the future. In order to make our tool replicable and sustainable, we share lessons learnt from our experience conducting SML, in the form of a set of recommendations for future analyses:

1. **Carefully construct evaluation questions, being conscious of which questions SML can and cannot answer.** Traditional impact evaluations using quantitative data are often concerned with causal questions, such as estimating the causal impact of an intervention on a set of outcomes. On the other hand, SML is better suited to address evaluation questions that are more descriptive in nature, and instead intends to capture the perspectives of stakeholders on a specific topic of online discussions. Examples of questions that SML can answer include: 1) Landscaping: What is the online debate in-country about a specific topic? ; 2) Actor analysis: What is the online influence of the stakeholder group of interest?; 3) Longitudinal study: How has the debate on a topic changed over time?

2. **Validation of data pulled from social media is key to ensure relevance.** The process of building search queries should be iterative and based on expert discussions. The set of key words need to take into account any linguistic or cultural sensitivities, while being simple enough to ensure a sizeable corpus. Once a search query is built, results should be analysed by experts and changed iteratively until a satisfactory level of relevance is reached. Note that this process can often be very sensitive – one word can skew the analysis significantly. Thus, it is recommended that multiple versions of the search query on a particular topic are ran.

3. **Exclusion and inclusion biases must be accounted for.** Depending on the evaluation context, inequalities in access to social media can result in certain groups being excluded from online discussions, often limiting the potential of studies using social media listening tools to be representative at the national, or even region level. The social media coverage in country, as well as factors potentially limiting access, should be considered to identify an appropriate stakeholder group whose online discussions will be analysed. Bias can also arise from the fact that people tend to be more extreme on social media than in person and are more likely to engage with controversial material. This further highlights the need for expert discussions to validate the relevance of online posts.

4. **External validity should be considered.** Given that online discussions are not necessarily representative of the entire population and in some cases the number of posts is small, the evaluation should clearly state whether the results can be generalisable. While an online debate on a particular topic is unlikely to be a true representation of an actual debate in-country, it is an interesting and complementary analysis that offers additional insights which would not be captured by traditional surveys.

5. **Meltwater vs self-programmed scrapers:** There are advantages and disadvantages to the use of online platforms relative to self-programmed scrapers, depending on the stakeholder group and the evaluation question at hand. While social media listening tools such as Meltwater cover a wider range of social media sources and thus allow to capture discussions of the general public, their analytical tools can be limited, for instance the inability to download raw social media data and carry out independent analysis. On the other hand, self-programmed scrapers allow us to access discussions of a specific stakeholder group, such as a Telegram chat for organisations working in the inclusive education space in Kazakhstan, and enable us to conduct our own Natural Language Processing analysis, but need to be written for every social media source separately.

6. **Use of alternative online SML platforms:** There might be platforms other than Meltwater with functionalities for scraping of social media data and analytical tools that are appropriate to use in your particular context. For instance, UNICEF could deploy the Talkwalker platform that is already available in-house. When choosing an online SML platform, the following factors are important to consider: 1) whether it is possible to download the raw corpus of data for independent analysis – Meltwater currently does not allow this; 2) the breadth of online sources that can be scraped, such as blogs, Instagram, Facebook, Twitter etc.; 3) the method for constructing search queries – in the case of Meltwater this is done through the use of Boolean operators, but that might differ depending on the platform; 4) analytical tools available; 5) availability of automated dashboards for displaying results; and 6) ease of access.

7. **Location analysis:** This has proven a tricky exercise in our experience, due to a large number of social media users choosing not to disclose the location of their post, possibly due to privacy considerations. Even in cases where the location is disclosed, it could be misleading due to the fact that major organisations and publishing houses tend to be located in large cities. This means that although the post is pinned to a particular location – its actual author could be located elsewhere. It also under-represents discussions from more remote locations that could have worse internet connectivity or physical infrastructure. Any location analysis undertaken using social media data should therefore be interpreted with caution.

8. **Demographic disaggregation:** As with location, identification of users and their demographic characteristics is often not possible. While raw social media data includes information on the content of the post, username, date, location (if disclosed), and related metrics – it does not disclose personal information of users. Therefore, the potential to carry out descriptive analysis disaggregated by demographic information is limited. However, some information about particular groups can be inferred by using specific key words in the search query. For instance, key words and their conjugations associated with girls or boys.

9. **Technical capacity:** It should be noted that replication of the tools described in this package, and in particular of the self-programmed scrapers, requires a high degree of technical capacity. The team should consist of at least one senior data scientist proficient with Python programming language and familiar with Natural Language Processing and text analytics. While the Python code we supply is annotated and accompanied by documentation describing the folder structure, each script, as well as expected outputs – it is not entirely automated. Certain paths are hard coded and

will need to be replaced in order to run on other machines. This means the code will require minor tweaks from the end user depending on the analytical context.

10. **Application to other countries and topics:** There is a degree of flexibility in our example search queries in Meltwater, as well as code for self-programmed scrapers to adapt it to a different country or topic of interest. In Meltwater, this would simply be done by replacing our “Location” domain with key words associated with the desired country. We recommend that in addition to the country name you also supply names of locations within the country, such as names of major cities. Similarly, the topic of interest could easily be replaced in Meltwater by creating sub-topics relevant to the discussion and combining them with the use of Boolean operators. In terms of self-programmed scrapers, since those were written to target specific organisations, they need to be adapted to the organisations or individuals that the user wishes to follow. This will require familiarity with the Python programming language.

Annex

Figure 6: Building a search query in Meltwater (Query 1: Inclusive Education KZ)

Figure 7: Building a search query in Meltwater (Query 2: Inclusive Education RU)

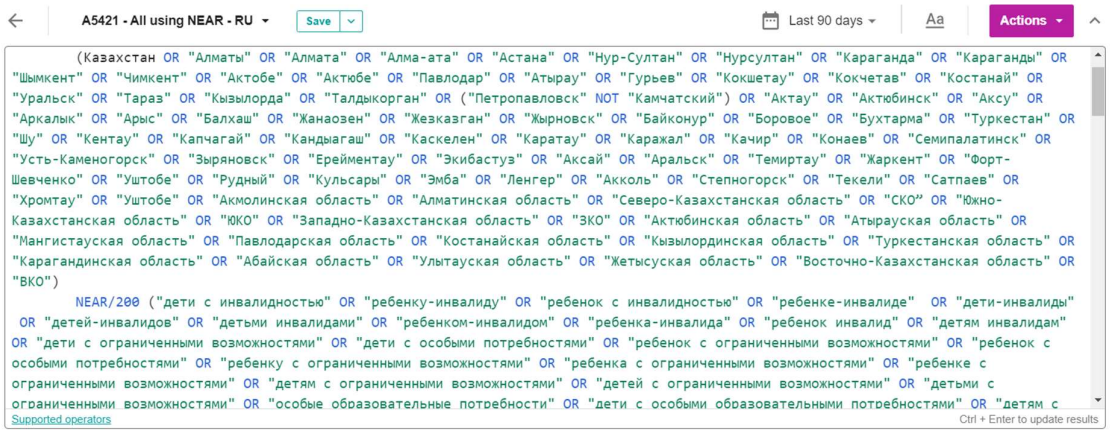


Table 3: Corpus 2, List of Organisations

WEB	TELEGRAM
https://bilimland.kz/ru/news-articles/news	Dara Charity Foundation
https://www.cpm.kz/ru/news	Mama Pro Kazakhstan кz
https://www.kzhol.kz/poslednie_novosti	Про Инклюзию | Инклюзия Жайлы (both group and chat)

https://orleu-edu.kz/ru/orleunews/	ПРАВА МЕНТАЛЬЩИКОВ	
https://special-edu.kz/news/6		
https://utemuratovfund.org/news		
https://bilimland.kz/ru/news-articles/news		
INSTAGRAM CHANNELS		
https://www.instagram.com/ardi_kz/	https://www.instagram.com/dara.inclusion/	https://www.instagram.com/pmpk_karaganda/
https://www.instagram.com/ario_kz_/	https://www.instagram.com/doskaz.kz/	https://www.instagram.com/prava_ocobennogo_rebenka/
https://www.instagram.com/asylmiras_astana/	https://www.instagram.com/inclusionteam/	https://www.instagram.com/ravmir_astana/
https://www.instagram.com/bolashakcharity/	https://www.instagram.com/juldyzai.fond/	https://www.instagram.com/specialedu.kz/
https://www.instagram.com/bulat_utemuratov_foundation/	https://www.instagram.com/kzhol_charityfund/	https://www.instagram.com/spectrum.astana/
https://www.instagram.com/centerbalama/	https://www.instagram.com/nuriya.baltabaikyzy/	https://www.instagram.com/veneraclub.atyrau/
https://www.instagram.com/centerkenes/	https://www.instagram.com/okoo.kz/	https://www.instagram.com/zeiin_atyrau/
https://www.instagram.com/centre.davinci/	https://www.instagram.com/orda_autism/	https://www.instagram.com/zhasurpaq_nur/
https://www.instagram.com/cpm_official_page/	https://www.instagram.com/orleu.edu.kz/	https://www.instagram.com/pmpk_karaganda/
https://www.instagram.com/dara.charity/	https://www.instagram.com/pmpk1_astana/	https://www.instagram.com/prava_ocobennogo_rebenka/