Russo-Ukrainian Conflict and Twitter: A Big Data Analysis

by Oscar Karlsson, Sam Hurenkamp, and Ramin Darudi

Abstract - Twitter, as an extremely popular social media platform, has for some time now found itself caught amidst global pressure to address the dichotomy within. Namely, it is caught serving as a medium for spreading news and global developments as they unfold whilst simultaneously serving as a playground for the machinations of organized misinformation and false news. The work within this report has performed analyses on a large public dataset using modern Big Data technologies to gauge public response to the ongoing Russo-Ukranian conflict. By having investigated the trending hashtags by frequency at a global level - this report serves as an introductory text towards further future statistical analysis.

Table of Contents

1. Introduction

The Russo-Ukrainian conflict began in 2014 (1) with Russia's annexation of Crimea, which later escalated in early 2022 when Russian forces crossed the border into Ukraine. Since then, the conflict has resulted in significant loss of human life and infrastructure in Ukraine, as well as suffering, economic disruption, and vast increases in political tensions worldwide.

In the ongoing conflict, social media tools such as Twitter are playing an important dichotomous role in not only spreading vital information about the conflict but also as a propaganda machine that spreads misinformation (2).

The works on which this report is built rely heavily on the successful implementation of popular Big Data tools and practices, the former of which is described in greater detail in Section 1.2 of this report. Although Big Data is not constrained by any one formal definition as of yet, Villars (2011) described it to be a measure of difficulty in overcoming challenges in assessing environments that experience rapid growth in data (3). De Mauro et al. (2015) later went on to admit that the rapid and unstructured rise of popularity in Big Data has resulted in the absence of a universally accepted formal definition (4), before proposing one as follows:

> *"Big Data represents the Information assets characterized by such a High Volume, Velocity, and Variety to require specific Technology and Analytical Methods for its transformation into Value."* (4)

Although the above-quoted definition does lend itself to realizing what constitutes Big Data as a concept, it does not provide a well-founded insight on what explicitly qualifies for classification under Big Data, as this is heavily reliant on current hardware restrictions. Data considered to be Big Data today, may not be so tomorrow. For that reason, and with our hardware restrictions in mind, it was assumed that any dataset over 10GB would provide a sufficient argument for consideration as a Big Data dataset.

This report aims to interpret global responses from users of the Twitter social media platform towards the ongoing conflict by assessing hashtag trends on the platform at a global level.

1.1 Dataset

The dataset chosen is a dataset sourced from Kaggle (5) and contains a myriad of tweets scraped from Twitter in the context of the Ukrainian war. The dataset is normally updated daily and is around 14 GB in size, which qualifies to be classified as a Big Data dataset under the aforementioned assumption. This dataset is a strong candidate for analytical work as it contains a lot of unclean data, for which the tools discussed next are well suited.

1.2 Introduction to the tools

1.2.1 Hadoop

Hadoop is an open-source framework under the Apache license which is used for storing and processing waste amounts of data using parallel computing. This is accomplished using HDFS (Hadoop Distributed file system) which allows Hadoop to splice up the data and distribute the workload to multiple computer clusters. The management of resource distribution is managed by YARN (Yet Another Resource Negotiator) which is used for scheduling the different tasks across the clusters.

1.2.2 Spark

Spark is an open-source data processing engine that is often used together with Hadoop for data analytics. Spark is a powerful tool that can be used for many different analytics projects because it has the ability to do a lot of different tasks in parallel on multiple computer clusters. Tasks that are commonly performed using Spark are; machine learning, graph processing, stream processing, and batch processing. Spark is written using the programming language Scala, but in this project, we are using PySpark. Pyspark is a python based API for interacting with Spark using Python.

2. Materials and Methods

This section describes the steps taken to give the different requirements of this project. Although the technologies employed allow for a multiple-computer remote cluster to perform the necessary actions and operations on the dataset, the work contained within this report was performed using one computer, the hardware specifications of which are found below.

2.1 Hardware

Hardware greatly limits the total data that may be processed on a system. As the main limiting factor of the works discussed, computer hardware plays a central role in pre-processing performances across the dataset. See Table 1 for the system hardware used.

Table 1. Overview of PC specs used in training, testing, and running the complex image classification model

CPU	AMD Ryzen 9 3900X (Stock Settings) - 12 cores/24 threads
RAM	G.Skill Trident Z 2x16GB (OC at 3600MHz) [F4-3600C16-16GTZNC]
Storage	Seagate 2TB HDD [ST2000DM001-1CH164]
Cooling	CPU: ARCTIC Liquid Freezer II 360 RGB [ACFRE00097A], top mounted in push-pull configuration.

2.2 Data collection

As the dataset itself was already scrapped, there was no need to do any further data collection. After initially installing and setting up Hadoop, all the CSV files of the dataset were written into the name node of the HDFS. As the name node and data node were hosted locally, subsequent read/write operations did not occur directly on the HDFS but rather Hadoop defaulted these operations to the local system for improved speed.

2.3 Data pre-processing

The data pre-processing phase involved reading the data from the HDFS to a PySpark data frame. In this way, the parallelization benefits of PySpark were realized over performing sequential operations using traditional Panda data frames, for which the dataset size was considered unsuited. Null values and dirty entries were removed to clean up the dataset and prepare it for analysis. Moreover, only the `userid` and `hashtags` columns were seen as relevant to the topic of investigation and subsequently the remaining columns were dropped from the data frame to save on space and performance. Henceforth, references to these columns of interest (e.g., `userid`) are made implicitly (i.e., userid) within the text. The hashtag column contained a dictionary of text and index values, where the former was the defining string and the latter described the hashtag text's position within the tweet body.

Hashtags row value contents were isolated and grouped by user ID before being aggregated to all other hashtags of that user by utilizing PySpark's `collect_list` function. In other words, this allowed all hashtag entries to be grouped under a distinct user id, whilst preserving all a user's hashtags values.

It is critical to note that Spark performs computation across two distinct techniques: Operation and Action. Operations, such as `filter` and `groupBy` are queued for future execution when an Action (e.g., `show`, `collect`, `count`) is triggered. This way, Spark optimizes future work under the hood, leading to more performant action execution. This mindset and methodology are then also applied to the PySpark API by definition and may lead to several memory or timeout errors as operations build up without an action being fired off. To circumvent this, and prevent these types of errors at the cost of performance, a `repartition` operation (to shuffle partition loads) and subsequent `show` action were performed frequently at the end of each major operation, such as filtering (which may lead to an unbalanced/skewed load across the select partitions).

2.3 Data Analysis, Representation, and Visualization

Analysis was made by means of filtering the dataset to draw a global choropleth map (see Figure 1), which served as a mask used to build a subsequent wordcloud (see Figure 2). However, as wordclouds make for pretty lazy visualization means (i.e., it is hard to extract precise information from it), horizontal bar graphs were also constructed but limited to a global top 10, 15, 20, 25, 30, 35, 40, 45 and 50 trending hashtags (an example of which can be seen in Figure 3).

Further analysis was intended for, and is discussed in greater detail in the Discussion of this report, but was ultimately discarded due to time constraints.

2.5 Security

Audit logging was enabled in the configuration XML files to provide an added layer of security to the existing works discussed in this report.

3. Results

Below are snapshots of the choropleth map for the world (see Figure 1), as well as an associated word cloud (see Figure 2) and horizontal bar graphs (see Figure 3). Further figures are available online, see Appendix B1 for the appropriate link.

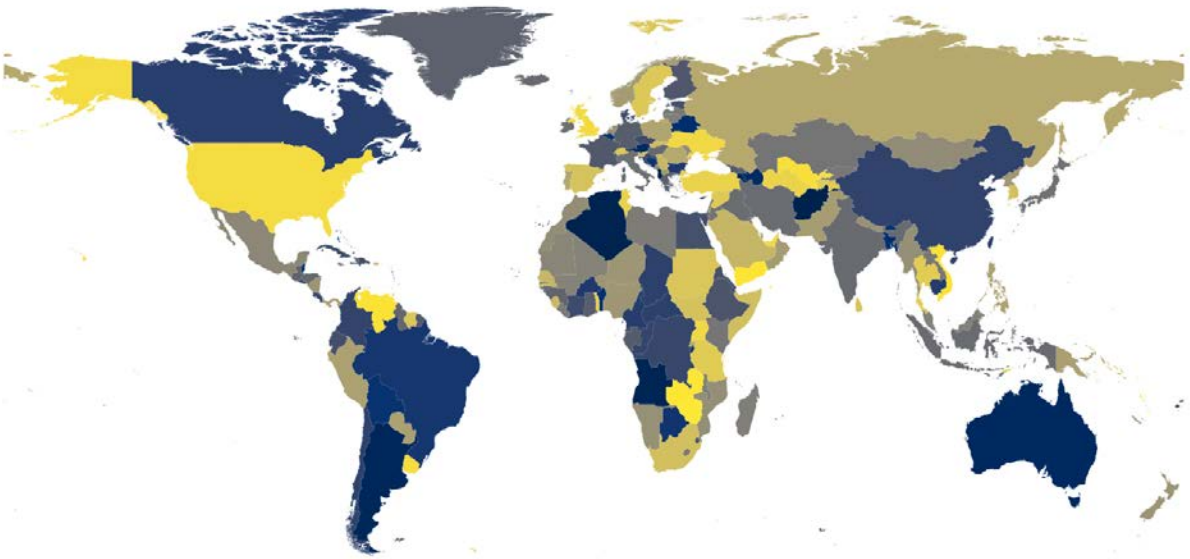

Figure 1: A choropleth map depicting the world and the boundaries of the countries therein, sourced using GeoPandas and a outsource SHP file (6).

Figure 2: A WordCloud that visualizes the most-trending hashtags globally by frequency, using Figure 1 as a mask.

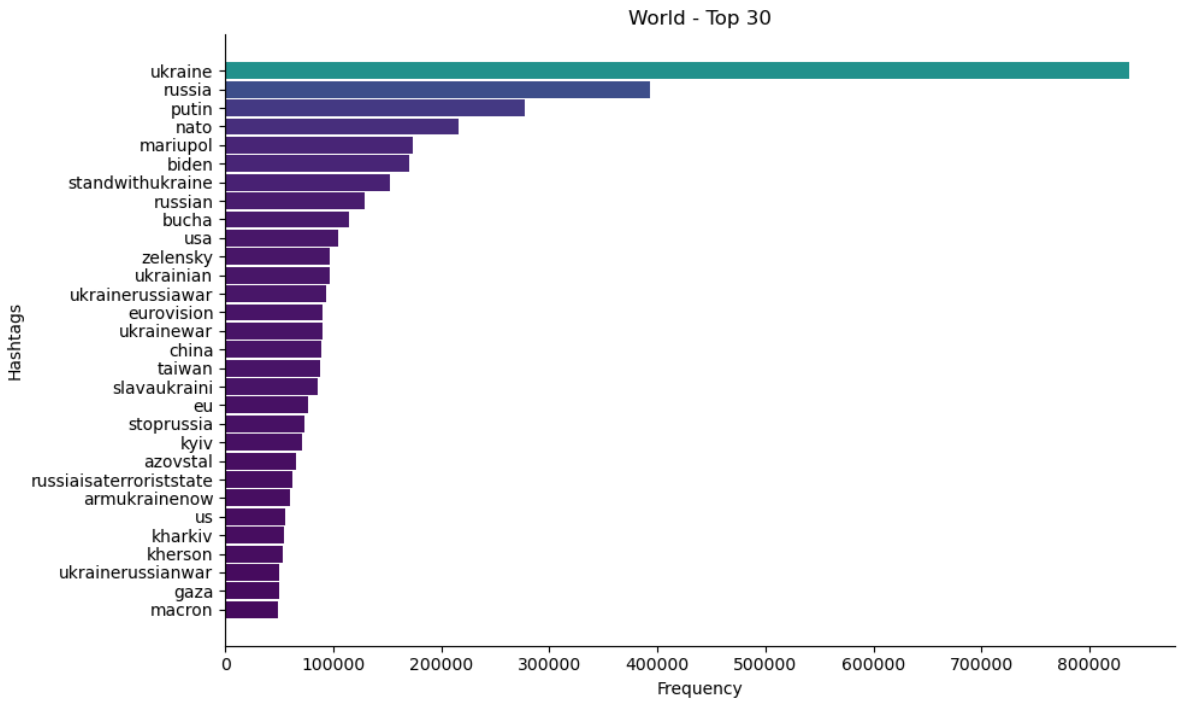

Figure 3: A horizontal bar graph that better ranks the top 30 global hashtags trends as seen in Figure 2.

4. Discussion

As no statistical model was employed to determine a measure of significance on the nature of the hashtags discovered during the data analysis, it is impossible to grade how well the results might have compared to a working hypothesis. However, it is clear from the Results Section of this report that the hashtags strongly associated with the current Russo-Ukrainian conflict are dominant within the data.

Further analysis of the nature of these tweets was intended and partially completed; such work included further analysis of trends per country but also to determine the probabilities on whether any one single hashtag was negative in sentiment, so that a trending sentiment per country could be uncovered. The results of these would then have been plotted on the world choropleth map depicted in Figure 1, through use of an appropriate gradient cmap. From this, further statistical analysis could have been made to determine if a significance existed within the data. However, it was not possible to complete these additional steps due to hardware limitations and time constraints. Still, it remains a strong candidate for future work.

5. Conclusion

The works enclosed within this report set out to investigate various Big Data tools and their effects on a large dataset (i.e., 14GB), as well as to perform further analysis, presentation and visualization into the current global hashtags trends, and were successful in doing so. This report provides both a useful insight into global response towards the ongoing Russo-Ukrainian conflict, but also a means to fuel further investigation using statistical methods to determine if there exists an underlying existence when, say, comparing sentiment across global regions.

The work carried out in this report was part of an ongoing tertiary education course project, the requirements for which are found in Appendix B - Table B1.

6. References

1. Dhawan M, Choudhary OP, Priyanka, Saied AA. Russo-Ukrainian war amid the COVID-19 pandemic: Global impact and containment strategy. Int J Surg. 2022 Jun;102:106675.https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows
2. Geissler D, Bär D, Pröllochs N, Feuerriegel S. Russian propaganda on social media during the 2022 invasion of Ukraine [Internet]. arXiv; 2022 [cited 2023 Jan 3]. Available from: http://arxiv.org/abs/2211.04154
3. Villars et al. - Big Data What It Is and Why You Should Care.pdf [Internet]. [cited 2023 Jan 3]. Available from: http://www.tracemyflows.com/uploads/big_data/idc_amd_big_data_whitepaper.pdf
4. De Mauro et al. - 2015 - What is big data A consensual definition and a re.pdf [Internet]. [cited 2023 Jan 3]. Available from: https://www.dhi.ac.uk/san/waysofbeing/data/data-crone-demauro-2015.pdf
5. UA Ukraine Conflict Twitter Dataset [Internet]. [cited 2023 Jan 3]. Available from: https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows
6. World Countries (Generalized) [Internet]. [cited 2023 Jan 4]. Available from: https://hub.arcgis.com/datasets/esri::world-countries-generalized

7. Appendix

Appendix A:

A1 - Project File Repository and Results:

https://github.com/sam-hur/DA381A-Project

Appendix B:

B1 - Project Requirements (table)

Table B1: This table shows the project requirements that need to be met in order to achieve a passing grade

Passing Grade (G) Requirements	**Implemented (Yes/No)**	**Comment**
Rq. 1: Data Collection	Yes	
Rq. 2: Data Preprocessing	Yes	
Rq. 3: Smart Data Analysis	Yes	
Rq. 4: Representation & Visualization	Yes	
Rq. 5: Security	Yes	added config settings for audit logging