

Abstract

This thesis explores the potential of applying machine learning (ML) to assist with open source intelligence (OSINT) analysis. As the shared input of both disciplines, data is the primary lens through which the topic is examined. To understand the entire process of deploying an ML model from data collection to analysis, an image classifier of Russian vehicle losses in the invasion of Ukraine was trained and tested. Trained on a dataset of over 50,000 labelled images from the WarSpotting database, the classifier achieved a decent accuracy of 79% on evaluation data on the five most populous categories of images. On testing data from a later period, the performance dropped to 62%. One explanation offered is that the static frontlines and the prominence of drones led to most of the recent imagery being aerial, while the training data was captured mainly from the ground. That result demonstrated how inevitable changes, even in seemingly well-curated data, can lead to the low performance of ML models in deployment. Beyond changes on the battlefield, deeper data issues came to light, including the cascading effects of early data management decisions and dataset imbalance. Overall, current image classification methods do not work well on the noisy data available.