Narcolands

A data driven approach to event based drugs popularity in the Netherlands

Ana Chaloska, Anisha Sloeserwij, Danny de Vries, Faye Tervoort, Jesse Kommandeur
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

KEYWORDS

drugs, google trends, court cases, Netherlands, visualization

1 INTRODUCTION

Every country is affected by crime, and so is the Netherlands. According to the Dutch center for crime prevention and safety statistics (CCV), over 90% of all organized crime is related to illegal drugs (CCV, 2022). The Netherlands owes this percentage to its prominent role in international drug trade (McDermott et al., 2021). Since problems surrounding drugs have increased considerably in recent years, authorities have prioritized and focused on identifying and tackling drugs related behavior (Eski and Buijt, 2016; Ferwerda et al., 2016).

In order to battle drugs related organized crime better, it is important for authorities, such as the police, to understand the modus operandi of criminal organizations through both academic and practical research. One of the organizations that focuses on the discovery of new modus operandi is the Police Academy (PA). In cooperation with the University of Amsterdam, the PA wants to focus on data driven research to identify drug related patterns in the Netherlands.

Previous research within the narcotics space focused on the supply-side of illegal drug trade (Paoli et al., 2013; Magliocca et al., 2019). This domain focuses on understanding the production, distribution, and trafficking of illegal drugs. It includes research on the modus operandi used by criminal organizations to produce and transport drugs, as well as the routes and networks they use to distribute them. This type of research is important for law enforcement agencies as it helps them identify and disrupt the operations of drug traffickers (McDermott et al., 2021).

Another area of research that is also present in prior literature is the demand-side of illegal drugs. This domain focuses on understanding the consumption and impact of illegal drugs on individuals, politics, and society (Flores-Macías & Zarkin, 2019; Riley, 2017). It includes research on the psychological and social factors that contribute to drug addiction, and the health and social consequences of drug use (Gonzalez, 2015; Sesnie et al., 2017).

These research domains are important for public health agencies and policymakers as it helps them understand the extent of the drug problem and develop effective strategies for addressing it. However, most of these studies are less relevant for authorities such as the PA since they focus less on quantitative data driven methods to identify patterns in society. Therefore, this study focuses on a data driven pattern identification of drug usage and events, such as festivals and public holidays. In order to investigate this relationship, this study aims to answer the research question: *"To what extent can we identify event-based drug popularity on online data resources?"*.

To answer this research question, the following sub questions were formulated:

- Which Dutch events are indicative for drug popularity?
- What drugs are used during events in the Netherlands?
- Which online data resources are relevant to identify drug popularity at events?
- How could we identify drug popularity at events?
- What are the differences between the specific events and drugs?
- How could we visualize event based drug popularity on online data resources?

The sub questions are relevant to the main research question as they provide a more detailed and specific understanding of the topic. The subquestion "Which Dutch events are indicative for drug popularity?" helps to narrow the scope of the research since it provides a starting point for identifying which events may be associated with drug use. Second, the subquestion "What drugs are used during events in the Netherlands?" is important for identifying which specific drugs are of interest for the research. "Which online data resources are relevant to identify drug popularity?" and "How could we identify drug popularity at events?" are crucial for understanding the methods and data sources of this research. "What are the differences between the specific events and drugs?" helps to identify any patterns or trends in drug popularity that may vary depending on the specific event or drug being considered. The last subquestion “How could we visualize event based drug popularity on online data resources?” is relevant since it allows us to transform our research data into a user-friendly dashboard to create insights for the PA.

In order to answer the research questions, we first present the theoretical state of affairs, followed by the methodological set-up, after which the results are presented. In the final sections, the most important findings are concluded and limitations are discussed, followed by recommendations for future work.

2 RELATED WORK

The collection and utilization of large amounts of data is a popular resource nowadays and is used in a variety of disciplines. [bron] The use of these large amounts of data, also known as big data, combined with the use of data analysis is having a major impact on the social sciences and humanities in general. [bron] Chan and Moses’ research shows that it has a particularly large impact for the specific field of criminology. [bron]

2.1 Predictive policing

Predictive policing is a term used for predicting certain behaviors or trends based on analyzing various data for law enforcement. [bron] Authorities use predictive policing to predict crime in order

to prevent the criminal activity instead of reacting to the crime that already occurred. For this reason, the domain and effects of predictive policing have been examined for a long time.

However, not all research agrees that predictive policing is a valid method. According to The New York Times' debate, predictive policing is a very effective way when it comes to predicting criminal behavior, but still contains many improvements in terms of ethicality. [bron] In contrast, Hardyns' research says that the analysis of such predictions has indeed shown its worthiness for different predictive systems in different areas, but exactly because it is a new development in the field of criminology, little is yet known whether it is an effective way for law enforcement. [bron]

Several studies have been conducted that have investigated whether Google trends can be a possible resource for predicting particular trends. Previous research has shown that Google Trends can be used as a predictor in different fields. Such as the healthcare industry, where Google Trends can help support the prediction of the outbreak of seasonal influenza and COVID-19 [bron], [bron]. Another study by Kassraie et al. used Google trends data in combination with Twitter data to predict the popularity vote of the 2016 presidential election in the US. They concluded that the combination of these social media platforms could be a mirror for the public opinion on political events [bron].

In the study by Perdue et al. it appears that Google trends can be a possible predictor for drug abuse trends. [bron] This is supported by the research of Gamma et al. which investigated whether there is a comparison between time trends of Google search interests and offenses committed in relation to the drug called Methamphetamine. From this study, it was found that law enforcement could indeed use the Google search feature as a possible predictor of Methamphetamine-related crimes. [bron]

3 METHODOLOGY

The approach to answering the research questions will be described in this section. The methodology consists of four parts. The first part describes the methods and literature that was used to select drugs and events. The second part describes the data pipeline which includes the data selection, data collection, data processing and data privacy. The third part describes the statistical metrics that were used to analyze the data. The last part covers the methods that were used to convey and visualize the data into the dashboard.

3.1 Methods for identifying drugs and events

3.1.1 Method for drug selection. To determine the different types of drugs which are used in the study, several institutes that collect data on drugs in the Netherlands were analyzed. The largest institution in the Netherlands that is related to drugs is the Trimbos Institute (TI). This national organization conducts research on the mental health of the Dutch people with a focus on the use of alcohol, tobacco and drugs. They involve all age groups in society and therefore cover the entire life cycle of citizens. This institute releases various analysis on alcohol, tobacco and other drug use in the Netherlands which includes reports on how to address these problems by naming prevention, education and policies.

In addition, The TI tracks drug-related developments through various monitoring systems. The most important monitor is the National Drug Monitor (NDM). The NDM collects and compiles all data on substance use, the drug market and drug-related crime of all ages in the population. This institution aims to provide a representation of the figures known in the Netherlands related to drugs. Using both data from TI and NSM, a set of specific drugs were selected for this study.

3.1.2 Method for event selection. First of all, it was determined which events would provide interesting results. Because of the subquestion "Which Dutch events are indicative for drug popularity?" it is interesting to use events where drug usage occurs a lot. Because where drugs are a popular means of enjoyment, they will also be used a lot. For this purpose, research was done on the use of drugs at different events. That way, the events served as a proxy where drug usage is known to increase. The first method that is used is literature research. This literary research was conducted to see if there has been previous research executed on drug use for different events in the Netherlands. However, very little is generally known about specific drug use per event in the Netherlands. Actually there is no research done in the Netherlands about drug use at specific parties or events. Due to the fact that it is difficult to investigate because few drug users want to be open about their drug use. For that reason we used another method where different news articles that say something about drug use per event are examined. This involves using different news sources, such as the AD, the NOS and the Telegraaf, as well as regional news sources such as LokaalGelderland and Echt Amsterdams nieuws. Using these two methods, a number of events have been identified where drugs are commonly used among the public. The research and answer to the sub-question is described in Chapter X.

3.2 Data Methods

3.2.1 Data selection. Before data could be collected and analyzed, a selection of data sources has been made. When it comes to the research question "To what extent can we identify event-based drug popularity on online data resources?", Google Trends data, Twitter data, and Dutch news data represent appropriate sources of information.

Google Trends data provides a comprehensive insight into the popularity of search terms and topics in real-time. By analyzing the frequency of drug-related searches on Google, researchers can obtain a thorough understanding of drug popularity across regions and over time (Batistic, 2021). Furthermore, Google Trends offers numerous filters and visualization options, making it a user-friendly tool for data analysis.

Twitter data represents a rich source of information on public perceptions and opinions (Bian, 2016). The platform is well known for its user-generated content and real-time information, making it a valuable resource for identifying the public's views on drugs. Additionally, Twitter's hashtag and trending topic features allow researchers to quickly identify the most popular drug-related topics on the platform.

Dutch news data is a crucial source of information in the assessment of event-based drug popularity. It provides a comprehensive understanding of media representation of drug-related events, including how these events are reported and framed, and the public's perception of drug use. By analyzing news data, researchers can

gain a deeper insight into the media's role in shaping public opinion on drug use (Mccombs, 2020).

In conclusion, the combination of Google Trends, Twitter data, and Dutch news data offers a diverse and comprehensive set of data sources for understanding event-based drug popularity on online data resources. These data sources provide valuable information on various aspects of drug use, such as search trends, public opinion, and media representation, making them a sound choice for answering the research question at hand.

3.2.2 Data collection. After data sources were selected, the data was collected. The Google Trends data was obtained using PyTrends, a library in Python for accessing and retrieving data from Google Trends (PyPI, 2023). This method of scraping allowed for an efficient and automated process for collecting Google Trends data, enabling the researcher to obtain a large amount of data in a short period. Only searches that were done on the territory of The Netherlands were considered.

Twitter data was collected using SNscrape, a tool for scraping social media data, including tweets and user profiles (GitHub, 2023). This method of scraping Twitter data provided access to a significant amount of real-time user-generated content, allowing us to research the popularity of conversations surrounding specific drugs in the netherlands. We based on a Twitter scrape query on the time interval January 2014 - December 2022 that contained the words 'xtc', 'cocaine' or 'ghb'. Since we are only interested in Dutch tweets, we excluded tweets that are non-Dutch. In total, the Twitter scraper collected more that 1.500.000 relevant tweets.

The Dutch news data was obtained by downloading a news corpus from the NOS, a Dutch public broadcaster. This method of data collection allowed us to access a large amount of news articles in a centralized Kaggle repository, which was last updated at the end of 2022 (Scheijen, 2022). In total, the NOS contained more that 250.000 relevant news articles.

3.2.3 Data processing. After the selected data was collected datasets were loaded into a dataframe for processing. The pre-processing of the obtained Google Trends, Twitter, and Dutch news data was crucial in ensuring that the data was in a format suitable for analysis. To this end, the following pre-processing methods were applied:

- Conversion of dates: All dates within the specified time frame were converted to the same datetime format. This standardization of the dates was essential for ensuring consistency in the data and making it easier to analyze.
- Calculation of week numbers: A function was created to calculate the week number of each date, as the time interval used for the analysis was "week." This function allowed for the grouping of data into weeks, making it easier to analyze trends and patterns over time. The data was aggregated on a weekly level because that was the most granular aggregation of one of the sources (Google Trends, 2023). To have a comparable analysis, the weekly aggregation was applied to the data from all sources.
- Topic feature extraction: A function was created to check whether a tweet, news article, or search query contained a specific word, such as "cocaine" or "xtc." This process of topic feature extraction was essential in identifying and isolating the data that was relevant to the research question and in understanding the prevalence and significance of specific topics in the data.
- Normalization: The extracted features were normalized, as the values were absolute, while the Google Trends data were relative. Normalization between 0 and 100 was performed, and this came in useful when performing statistical tests and visualizing the data in the dashboard. Normalizing the data allowed for comparison and analysis of data from different sources, as the data was all expressed in the same unit.

Overall, the pre-processing of the Google Trends, Twitter, and Dutch news data was essential in ensuring that the data was in a suitable format for analysis. The conversion of dates, calculation of week numbers, topic feature extraction, and normalization were critical in making the data easier to analyze and interpret, and in ensuring that the results obtained from the analysis were accurate and reliable.

In addition,other pre-processing techniques such as sentiment analysis, locational feature extraction, word2vec synonym extraction and cleaning function have also been created. These functions could be used in future work to extend the research, make it more precise or compute detailed micro-level information.

3.2.4 Data privacy. In conducting the data collection and preprocessing of event-based drug popularity on online data resources, data privacy was a significant consideration. All data sources used were subjected to ethical and privacy considerations to ensure that all personal information was protected and that the data was collected, processed, and used in a responsible and ethical manner.

It was ensured that all data sources used were publicly available and did not contain any sensitive or personal information. The Google trends data was collected via Pytrends, in which personal information is already anonymized. For example, Google Trends data did not contain any information on the users who did the searches, but only the location where the search happened. The Twitter data was obtained using the SNscraper, which only collected data from public profiles and ensured that the data collected did not contain any personal information. For example, username and tags were removed.

In addition, appropriate measures were taken to protect the privacy of individuals and organizations that provided the data. For instance, all data was de-identified to remove any personal information that could be used to identify individuals or organizations.

3.3 Statistical methods

3.3.1 Hypothesis. To identify event-based drug popularity at events the following hypothesis was formulated: There is a significant increase in the metrics (number of searches, -tweets and -news articles) of each data source during the weeks of the events known for increased drug usage.

3.3.2 Computing peaks per regular week. A metric was constructed in order to translate the hypothesis into a test design. The data of each data source was aggregated weekly to represent the counts per searches, tweets and news articles respectively.

The aggregated data was ordered chronologically and plotted on a line plot. The peaks in the line plot were found using the

method signal.findpeaks from Pythons library Scipy [reference]. This step provided evidence on the weeks where there is an increase in the representative metric in relation to the other weeks in the considered time range (2014-2022).

The data with labeled peaks was stored in a dataframe in the form of a Bool column coupled with a column for the corresponding week number. For instance if there was a peak in Week 2 2022, the Bool column would be populated with True, else it would be populated with False. With that each week number was represented with 9 rows, 1 for each year between 2014 and 2022. Finally, a proportion was calculated for each week representing the percentage of rows with "True" value out of the 9 rows for each week. With that, each week was represented with a number representing the percentage of years where a peak was observed for that particular week [Table with example from slides - Appendix].

3.3.3 Computing peaks per event week. The above steps were performed again in order to represent the percentage of peaks for the weeks of the events with increased drug usage (King's day, ADE, Pride, Lowlands and New Year's Eve). This was needed because most of these events do not always fall in the same week of the year.

For every year the dates of the events were identified. This was necessary because some of the events happen on a different date every year (e.g. ADE, Pride) [Dates per event - Appendix]. Based on the event dates, the week in which each event happened were identified. [Appendix]. Based on the data extracted from the line plot labeled with peaks, the years in which a peak happened per week of event were identified. Similarly as in the peaks per week, the percentage of weeks in which a peak was observed were calculated.

3.3.4 Test design. A p-value approach proportion test [reference] was performed on the generated data in order to test the hypothesis above. The test was once for each combination of drug, event and data source. Below are the specifications of the test:

- Null hypothesis (H0): The proportion of peaks observed in drug related event weeks is not significantly larger than the proportion of peaks in regular weeks.
- Alternative hypothesis (H1): The proportion of peaks observed in drug related event weeks is significantly larger than the proportion of peaks in regular weeks.
- Level of significance: To determine level of significance, an alpha of 0.05 was used. This implies 5% risk of concluding that there is a larger proportion when there in reality that is not the case.
- Test statistic (z): The test statistic was calculated based on the sample proportion and population proportion [formula - Appendix], which is suitable for a proportion test.
- Accepting the alternative hypothesis: The p-value was calculated from the test statistic. A p-value of less than 0.05 (corresponding to the significance level) indicated evidence to reject the null hypothesis and accept the alternative hypothesis, and hence prove that the proportion of peaks during event weeks is indeed larger than regular weeks.

4 REQUIREMENTS

These user requirements are derived from the case description provided by the client as well as feedback from the stakeholder during the initial ideation and prototyping phase of the project and where further narrowed down during the project design workshops using the MoSCoW priorization method. The term 'user' in the requirements refers to two specific types of similar target audiences that will make use of our prototype, police agents who want to explore the dataset to gain insights and data analysts who want to filter and compare our datasets.

(1) (M) The user must be able to use the prototype on a personal computer and interface with a screen
(2) (M) The user must be able to filter the datasets to compare different years of data
(3) (M) The user must be able to filter the datasets to compare different types of drugs
(4) (M) The user must be able to overlay multiple datasets and trend lines on top of eachother
(5) (M) The system uses open-source software and not locked-in corporate data tools
(6) (M) The system has a user-friendly visual design and interaction design
(7) (S) The user should create an account to store specific and personalized filters
(8) (S) The user should be able to download the raw datasets in specific file formats
(9) (S) The user should be able to navigate between different overviews showing corresponding data
(10) (C) The user could upload their own dataset and sources of specfic drugs and news sources
(11) (C) The system uses real-time up-to-date API data

5 RESULTS

5.0.1 Prototype (dashboard).

Based on these requirements and the datasets a custom web-based visualization was created to allow the analyst to visualize explore, compare and filter the datasets. When filtering the charts update in real-time with keyframed animations. Screenshots and a live version of the prototype included in appendix 2a. For this prototype demo the user is logged-in with a default user account which is shown in the sidebar navigation. From there the user has four different overview pages to navigate to.

Events overview page

The first is an events overview page which shows a line chart with on the x-axis the week numbers per year and a relative score from 0 to 100 on the y-axis for our three datasets, the Google Trends, NOS News data and Twitter tweets. On the left bottom is a legend with a list of drug-related events. The user can click on an event which will highlight the week in which the event occured in the line chart to more clearly visualize peaks in the data which indicate event-based drug popularity. A small description of the event, most popular drug and event dates also is shown. The bottom-right shows the Google Trends data in a relative score over all 5 years of our dataset. This shows more clearly shows the lack of peaks of events in covid pandemic years.

Region overview page
The second page is a chloropleth map page showing the regions of the Netherlands. Each region is color coded using a linear color scale based (blue interpolation) on the relative score of Google searches for that region. On this page the user also has the option to focus on years and drugs but also on specific event dates. This for example shows that in Flevoland when the Lowlands event is happening there are alot of searches in that region for the drugs XTC.

Related queries page
The third page is the related queries page which shows a line chart for our three chosen drugs as filter options. The y-axis is a relative score from 0 to 100 and the x-axis are the week numbers for our specific year. The user can filter between different years. On the bottom are three polar charts which show the related drugs people are searching for with a relative score and a subset of the drugs in the legend. For example, people who search for cocaine also search for crack very often. It also shows that before 2019 GHB was not searched for by many people. It increased in popularity after that year.

Settings page
A settings page is also included. This page shows the different datasets used in the dashboard and a download button which allows the user to download a specific dataset either as .csv or .json to store on their computer.

5.0.2 Web application. The prototype is a web-based application using web standards and open-source software and libraries. It being web-based allow the dashboard to run operating system independent. Only a webbrowser installed on the users device is required. The application will mainly be used in a desktop environment by the analyst so the dashboard is not fully responsive and not mobile optimized. The web application is created with the open-source front-end framework Svelte [1] and UI framework SvelteKit which allows the application to be build in components, each chart is rendered seperately making it more efficient to add functionality (add datasets, render different chart types) in the future but also makes the dashboard performant when more data and charts are added since Svelte already pre-renders the page offloading work from the browser. Working in components and with a frameworks such as Svelte allows future web developers to get up and running fast and add more functionality in a progressively enhanced manner. Svelte can be downloaded as a module (package) from NPM [2] and uses the JavaScript back-end run-time Node.js [3].

For the charts the JavaScript charting library Chart.js [4] is integrated into the components which allows charts to be rendered in HTML5 Canvas without much configuration. With Chart.js you can add a specific dataset and the scales of the axis will automatically change accordingly to the scale defined by the data. The filter options and updating of the charts is more custom, it uses JavaScript utility functions to allow the data of the to be preprocessed and have only the data changed not the scales of the whole charts. For the map page an additional Chart.js Geo Plugin is used to render the Chloropleth map. It uses a TopoJSON file to render the regions of the Netherlands and filters the properties within that .json file to render a score for each of the region.

The source code for the dashboard is published open-source on GitHub using the MIT license. A live version of the dashboard is contionously deployed on hosting platform Netlify [5]. Corresponding links can be found in appendix 2b.

6 CONCLUSION

6.1 Give an answer to the research question

7 FUTURE WORK

7.0.1 Expand data sources. A subset of drugs and years was used for our research. To allow for more exploration and filterability the dashboard could be expanded with more datasets. Especially more relevant drugs (upcoming drugs, non-legal substances, NSPs) could be added as filter options. We also currently only use NOS news data but other more popular dutch news sources (e.g. RTL nieuws, Nu.nl) could be scraped to give a more accurate representation of the popularity of drugs mentioned in news articles. This will both improve the quality of the dataset by more accurately calculating a relative score as well as quantitify by gathering multiple sources and aggregating.

7.0.2 Real-time API data. Currently the dasboard relies on exported data that is then loaded into the web visualization. As a futher enhancement for the prototype and to make it more dynamic is to have the dataset exposed as an API which the dashboard can then fetch getting up-to-date real-time data. In this beta version datasets need to be added manually to the GitHub repository.

7.0.3 Usability testing. Further usability testing needs to be done to validate the User Experience (UX) and User Interface (UI) of the prototype to more accurately represent the needs of the police analyst. Basic user testing was done on a small group of people and feedback from the client was incorporated. But still, a lot of assumptions about the user have been made. User testing the live version of the prototype to a larger user base will uncover hidden problems. The feedback from the users would further validate the workings of the prototype.

In general further research is required to conclude that online data-gathering tools fare good indicators for predicting the popularity of specfic drugs.

8 DISCUSSION

9 APPENDIX

[1] https://svelte.dev
[2] https://www.npmjs.com
[3] https://nodejs.org/
[4] https://www.chartjs.org
[5] https://www.netlify.com