Maaß, Christina H.

Conference Paper

Can bad news be good predictors? Illuminating the dark figure of crime with crime-related news

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage"

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Maaß, Christina H. (2023) : Can bad news be good predictors? Illuminating the dark figure of crime with crime-related news, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage", ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/277607

Can bad news be good predictors?
Illuminating the dark figure of crime with crime-related news

Maaß, Christina Heike
University of Hamburg, Hamburg, Germany
christina.maass@uni-hamburg.de

Version: 4th September 2023, 9,659 words

Abstract

Unreported crimes pose a threat to economies and societies worldwide as they prevent state authorities from effectively addressing crimes. Yet the only (incomplete) measure available are victimization surveys. This paper sheds light into the dark of unregistered incidents by investigating the informational value of a new data source, crime-related news articles, in a machine-learning context. Centre of the approach is a text analysis of news reports augmented by macroeconomic variables and monthly dummies. With this approach, we provide a new tool to approximate overall crime levels in the United States as indicated by the National Crime Victimization Survey timely and with high accuracy. Our approach enables improvements in resource allocation, increased public safety and thus greater economic prosperity.

JEL classification codes: E26, K42, C40

1. Introduction

Crimes have been a threat to peaceful coexistence in society since the beginning of humankind. Despite increased efforts to prosecute crimes today, official crime statistics do not cover all committed crimes. Federal criminal offices indicate that factors like general reporting patterns, the scope of police controls, changes in statistical recording or criminal law, and real changes in the occurrence of crime affect the extent to which crimes are captured in these statistics (BKA 2022). The main factor, however, seems to be simple non-reporting by citizens as already found by Skogan (1977) and confirmed by low reporting rates in recent victimization surveys. Therefore, unreported crimes are the focus of this paper.

There are various reasons why crimes are not reported to the police. From a political science perspective, the decision to report a crime is based on a cost-benefit analysis that includes, for example, an assessment of the seriousness of a crime, police efficiency, demographics, and the relationship between victim and offender (Skogan 1984). Approaching reporting of burglaries from an economic point of view, MacDonald (2001) shows that individual willingness to report varies over time and with the economic cycle. Thus, factors that are directly related to economic prosperity like

employment status or the amount of financial loss involved play an important role (MacDonald 2001). Additionally, possible insurance claims, the perceived obligation to report, attitudes towards the police, feelings of being guilty themselves, the reporting behaviour of third parties, and self-help behaviour shape the dark figure (Skogan 1984; MacDonald 2001). Criminological research reveals differences between neighbourhoods depending on their cohesion and socio-economic status (Goudriaan et al. 2006). Sociologists point out that non-reported crime should not be equated with non-resolved crime as victims might decide to handle the issue by themselves (Kennedy 1988), which possibly leads to even more criminality. Due to all these factors, crime is per definition a phenomenon with a considerable number of incidents that are not recorded by the police. This number is frequently referred to as the dark figure of crime (Coleman and Moynihan 1996).

To know as much as possible about the reality of criminal offences is of high relevance for society and decision-makers (e.g. in the department of justice or in police departments) for a complete assessment of the situation and thus for adequate prevention and intervention measures (BKA 2022). A substantial dark figure is dangerous for an economy, as the limited picture of the current crime situation makes it difficult for state authorities to fight crime efficiently, and for a society, as individual uncertainty regarding one's own security promotes distrust of state institutions and strangers. Despite the longstanding interest in overall crime, the only measure of unreported crimes policy makers have to rely on are victimization surveys. While they provide valuable insights, they are costly, only available with a delay, and rarely challenged by other estimates. To meet the demand for new approaches to measuring actual crime rates, we investigate a new pillar in research on the dark figure of crime: the use of news reports in a machine learning context to predict the development of the monthly number of incidents as indicated by the Victimization Survey in the United States of America. We augment more traditional information from the economic indicators monthly gross domestic product (GDP), unemployment rate and inflation rate as well as from month dummies to control for seasonal variation with novel information from news articles in form of text data. We contribute to getting closer to the true number of crimes by proposing a new approach that sheds light into the dark up to two years earlier compared to surveys at lower cost. This is of high practical relevance, because economic costs arising from misallocations and security concerns can be reduced with our model. In other words, applying our approach enables improved resource allocation, increased public safety and thus greater economic well-being and welfare.

The hypothesis underlying this research is that bad news (those that are related to crime) contains hidden information on the overall occurrence of offences and can thus be a good predictor of actual crime levels in a country. Although we do not expect to capture all committed crimes with this

approach, it provides an additional point-of-view that can be used in conjunction with existing methods to better understand patterns in the overall number of criminal incidents. In this study we aim at investigating the overall number of crimes that also includes unregistered ones and to thus enrich the research on the part of the number that cannot be measured. Questions related to the exact number of committed crimes in a given period as well as discussions tackling definitions of crime and differences depending on types of crime, regions etc. are not part of this paper.

The paper is structured as follows: Section 2 explains the relevance of knowing the actual number of criminal incidents, section 3 describes approaches to measure criminal incidents. Section 4 is concerned with theoretical considerations. Section 5 describes the data and method. Section 6 covers the results and section 7 the discussion and conclusions.

2. Relevance to investigate the actual number of criminal incidents

To know as much as possible about the true extent of criminal incidents is crucial as crimes are a huge problem for society and social coexistence. For individuals, a high share of crimes being registered (and possibly resolved) increases the feeling of safety and it allows victims to enforce their benefit claims (Skogan 1977, 1984). From a socio-economic and institutional economics point of view, a low dark figure would indicate an efficient constitutional state and a coinciding jurisdiction and legal effect. Consequently, when a society believes in the rule of law and the constitutional state functions, the population internalizes the norms. In other words, when the dark figure is generally low in a society, its members have the intrinsic motivation to behave correctly. Additionally, effective crime combating is essential for a democracy. The separation of powers is a relevant pillar of democracy. The police are part of the executive in the separation of powers. The media are part of an additional, virtual pillar in the separation of powers, the so-called fourth estate, which exerts a strong controlling influence on political events. The interplay between information from the media and the police opens up new possibilities. This is of high relevance for the police as they need to know as many essential aspects of the reality of crime as possible for complete assessment of the situation, to legitimize itself and to improve public perceptions of the police. Furthermore, it is crucial to justify the budget, to allow for an optimal allocation of budget and police patrols, as well as for adequate prevention and intervention measures (Skogan 1977, 1984; BKA 2022). This approach can be seen as additional support for the police (and society), that can make use of new data sources in addition to existing tools.

Uncovering unreported crime thereby is relevant for most crime types, rather traditional ones like burglary and theft as well as recently more consciously perceived ones like hate crime (Myers and

Lantz 2020; Pezzella et al. 2019). In this paper, we do not distinguish between different types of crime.

3. Procedures to measure the number of criminal incidents

There are two official procedures to measure the occurrence of crime in the US guided by the Department of Justice. They measure a distinct group of incidents and thus deliver different estimates of crime occurrences as the methods and definitions of the two approaches differ to some extent (Morgan and Thompson 2022; Morgan and Smith 2022). Firstly, there is the Uniform Crime Report (UCR) which includes the National Incident-Based Reporting System (NIBRS) and comprises incidents registered by law enforcement agencies throughout the country. This represents the official police statistics and thus does not contain information on incidents not reported to the police. Secondly, there is the National Crime Victimization Survey (NCVS, formerly called National Crime Survey), which consists of reported and unreported incidents individuals name when surveyed. This type of survey is part of one of the first approaches to quantify the dark figure of crime, which emerged in the 1960's (Coleman and Moynihan 1996).[1] These surveys provide an additional measure of crime, namely of crime known to the public instead of crime reported to the police (Coleman and Moynihan 1996). The idea behind this is that surveying a random sample of the population would help to detect victims and thus provide helpful insights into the true prevalence of crime (Biderman and Reiss 1967). The advantage of victimization surveys is that reporting an incident to them is anonymous and there are no consequences, both of which relate to reasons for not reporting to the police (Biderman and Reiss 1967). Studies of victimization around the world reveal that there is a substantial dark figure of crime in every territory (Skogan 1984). Nevertheless, they are subject to different methodological issues like the necessity that a victim is aware of its victimization, difficulties in collecting a representative sample and non-random measurement errors (Coleman and Moynihan 1996). For a comparison of the measurement of crime by police statistics and victimization surveys see e.g. Biderman and Lynch (1991) for a description from the 1990s and Ariel and Bland (2019) for a description for Great Britain. Another survey-based approach are so-called offender surveys questioning possible offenders whether they have committed a crime. However, as they are expected to be little reliable, they are not part of the analysis in this paper.

[1] A detailed description of the developments prior to the introduction of victimization surveys, especially in England, can be found in Castelbajac (2014).

The significance to know as much as possible about the true occurrence of crime encouraged scientists to approach the dark figure of crime since the late 1960s, with Biderman and Reiss (1967) being one of the first proponents to approximate it by evaluating surveys (Castelbajac 2014). They found that especially minor incidents are subject to non-reporting (Skogan 1977). It is shown that augmenting incidents recorded by the police with survey data allows for area-specific estimates of the dark figure of crime and can help reduce measurement errors in police statistics (Buil-Gil et al. 2021).

Whereas research based on traditional methods often investigates crime reporting from a microeconomic point of view, more recent approaches also allow for macroeconomic analyses. For example, MacDonald (2001) finds, that the dark figure of crime follows the economic cycle (MacDonald 2001). Another macro-oriented approach towards a better understanding of crime patterns can be found in the area of predictive policing (see e.g. Kaufmann et al. (2019)). In this discipline, individual behaviour is targeted by feeding an algorithm with information regarding time, location, neighbourhood characteristics or traffic infrastructure of past crimes. Based on the pattern found in the data, future crimes can be foreseen, which gives authorities the opportunity to act early (Kaufmann et al. 2019). Further approaches taken by the US-police include predictive hotspot mapping and risk terrain modelling (Babuta 2017).

Current research indicates that it is imperative to adjust criminological research to the new circumstances in a digitized world and to apply new methodological approaches (Smith et al. 2017). In the last ten to twenty years, new strands of literature dealing with various approaches emerged. Scientists augment calls for service data with spatial video data and crime perceptions of (ex-)police officers and community members to analyse spatial differences in crime reporting within a neighbourhood (Porter et al. 2020). In order to measure trends in violent crime, the number of hospital admissions that result from violent acts can be used (Estrada 2006). Making use of big data, it turns out that Twitter posts are related to fear of crime in a population (Curiel et al. 2020) and that tweets indicating neighbourhood degeneration can serve as a proxy for police-recorded crimes in low-crime areas (Williams et al. 2017). Using statistical language processing and spatial modelling, it is shown that the content of Tweets can improve predictions of the occurrence of different types of crime compared to a standard estimation approach (Gerber 2014). Augmenting traditional demographic and geographic crime data with Point-Of-Interest and taxi flow data is found to improve predictions of neighbourhood crime rates in the city (Wang et al. 2016). An overview of further significant contributions can be found in Oatley (2022) regarding big data applications for crime analytics and in Shah et al. (2021) regarding the use of machine learning and computer vision

for crime prediction. Despite increased awareness for the dark figure of crime in the last decades, the only measure for unregistered incidents to date is victimization surveys. In particular, to the best of our knowledge, there are no papers concerned with predicting trends in the actual number of crimes. In order to close this gap, this paper uses a new data source together with machine learning to show the possibility for predictions that include unreported crimes based on news articles.

4. Theoretical considerations

As unreported crimes can pose a threat to economies and society in general and to economic and societal well-being in particular, the topic is of high relevance for economists. In the classical sense, the scholar economics of crime is especially concerned with normative considerations about the optimal allocation of resources towards punishment of offenders, economic incentives for criminals, cost-benefit analyses and viewing crime from a market perspective (see e.g. Becker (1968), Ehrlich (1973), Freeman (1999), Chalfin and McCrary (2017), Draca and Machin (2015)).

In our paper, we chose a different approach and estimate the total number of incidents in a month as reported by the NCVS focusing on a methodology that allows us to incorporate a new data source available at high frequency: text data from crime-related news articles. We show that this data contains valuable information for predictions of the true level of crime in a country. Following the recommendations by Williams et al. (2017) and Kaufmann et al. (2019) for criminological research with big data, we make sure to have a theoretical foundation for the use of our big data source as well as to use the text data in conjunction with traditional indicators to avoid spurious regression results. The theory behind using crime-related news articles is firstly based on the work of Porter et al. (2020), who show that crime perception of community members like police officers helps to detect spatial differences in crime reporting. Following this approach, we use crime perception of journalists expressed in news articles. This measure could serve as a proxy for aggregated crime perception in a country. Secondly, it appears reasonable that some victims might shun the bureaucratic effort to report a crime, but might be willing to tell a journalist about his or her experiences. Thirdly, different papers showing informational value from twitter posts (as explained above) make us confident to believe that news articles should be at least as reliable as random tweets from individuals. As traditional data for the analysis we use the monthly GDP and the unemployment rate, following the intuition by MacDonald (2001) that the economic cycle as well as the employment status have an effect on reporting behaviour. To complement the most relevant economic indicators, we also incorporate inflation. Additionally, we include month dummies to control for seasonal variation.

The aim of our analysis is to explain developments in past months using information from the respective months. Applying our approach to the present would allow to predict the development of the number of incidents in the preceding month.

5. Data and method

As an approximation for the dark figure of crime, we use the total number of incidents (including unreported ones) from the NCVS in the USA. This inquiry is a rotating panel survey conducted biannually by the United States Department of Commerce. U.S. Census Bureau since 1973 (U.S. BJS 2021b; NACJD 2022). In each survey, around 49,000 households (approximately 100,000 individuals) are surveyed (NACJD 2022). In addition to investigating unreported crimes, its main objective is to provide information on victims and the consequences of crime, and to measure the different types of crime (NACJD 2022). Study participants from the age of 12 are asked whether they have experienced some kind of offense (e.g. burglary, robbery, assault, theft) in the past six months. The survey collects detailed information on up to three incidents and asks for demographic characteristics of the respondents. For our analysis we focus on the surveys from 2016 (U.S. BJS 2020a), 2017 (U.S. BJS 2020b), 2018 (U.S. BJS 2020c), 2019 (U.S. BJS 2020d) and 2020 (U.S. BJS 2021a). With these four datasets we are able to analyse the period from January 2016 until December 2019. Our target variable consists of the difference between the incidents in the current month and those in the previous months. This difference is binned into 4 bins of equal size ranging from -450,000 to -225,000, from -225,000 to 0, from 0 to 225,000 and from 225,000 to 450,000.

There are three different classes of features used in the empirical analysis to explain the variation in the actual figure of crime. The first one consists of monthly values for the macroeconomic indicators GDP, unemployment and inflation. The data for the economic variables stems from two different sources. The US monthly unemployment rate is obtained from the US Bureau of Labour Statistics (U.S. BLS 2022) and is used as the difference in percentage points compared to the previous month. Monthly GDP as normalized values, used as the monthly growth rate in percent, as well as US monthly inflation in form of the development of the consumer price index (CPI) in percentage points are obtained from the OECD (OECD 2022b, 2022a) for the years 2016 until 2019. The macroeconomic variables are scaled using the Min-Max Scaler to obtain non-negative values. The second class of features consists of twelve dummies for each month of the year. These first two classes serve as baseline features to guarantee reasonable results when using the news articles.

The third class is a big data source consisting of text data: the Dow Jones Newswire (DJN)[2] reports published by the US-American publishing house Dow Jones & Company from 2016 until 2019. These reports contain economic and financial news in real-time that stem from 46 different news services like Dow Jones Institutional News, The Wall Street Journal or Dow Jones Energy Service. For our analysis we focus only on crime-related news from the US (both as categorized by the newswire itself) and drop duplicates, which leaves us with a total of 36,974 news articles. After erasing empty phrases at the end of each report, we change all words to lowercase, delete punctuation, special characters as well as numbers and reduce them to their stems using the Snowball stemmer. The news articles are converted to a machine learning readable format using Term Frequency – Inverse Document Frequency, which indicates how representative each word is for the respective document. We only keep words with at least two letters that occur at least 10 times in the corpus and remove stopwords. We are then left with 12,500 words in the aggregated analysis and 17,099 words in the high frequency analysis that are going to be used as features.

We visualize the most frequent words from the DJN dataset in a word cloud and based on Latent Dirichlet Allocation (LDA) (Blei et al. 2003) we gain insights into different topics present in the news articles. The centre of our analysis is a retrospective forecast of the actual number of criminal incidents. It is retrospective due to the fact that we explain the developments in a past month using available data for that month, that would not have been available a month before. This is a substantial advancement, since information from the surveys is only available with several months of delay. Survey results are published in autumn of the following year, so that for analyses of the second six months of a specific year, interested persons need to wait around 2 years, as part of the data for these months is included in the survey of the following year. For the empirical analysis we use a machine learning approach, which allows us to extract information from the highly frequent and unstructured news data. The method chosen is Complement Naïve Bayes (CNB) due to its high performance when using text data combined with speed and easy implementation. This method belongs to the class of Naïve Bayes methods that use Bayes theorem to classify data using the "naïve" assumption of independence between the features given the label of the target class (Zhang 2004). The CNB is an adjustment of Multinomial Naïve Bayes allowing for skewed training data, which increases effectiveness and reduces the bias inherent to skewed data (Rennie et al. 2003). Following Rennie et al. (2003), the CNB estimate for the probability that word *i* occurs in any class except *c* is shown in Eq. (1).

[2] https://www.dowjones.com/professional/newswires/

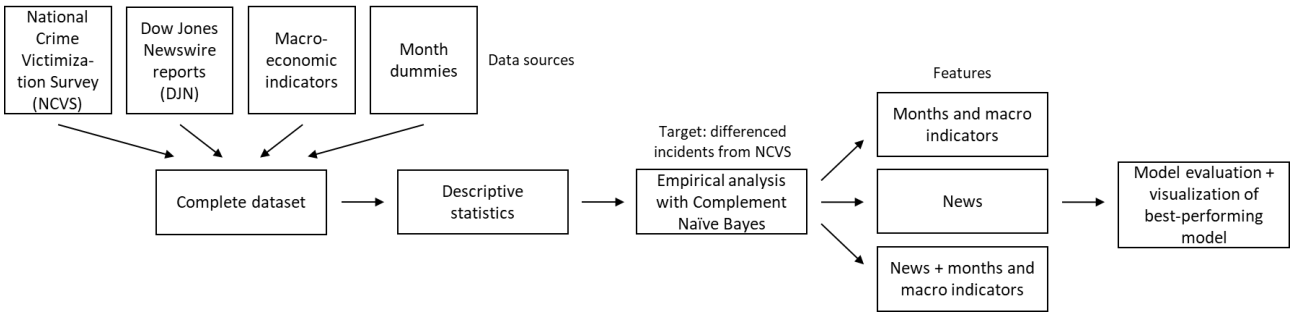
$$\hat{\theta}_{\tilde{c}i} = \frac{N_{\tilde{c}i} + \alpha_i}{N_{\tilde{c}} + \alpha} \tag{1}$$

Thereby, $N_{\tilde{c}i}$ indicates the frequency with which word *i* occurred in documents in classes different from c, $N_{\tilde{c}}$ indicates the overall number of word occurrences in classes other than c. α_i represents smoothing parameters, adding in imagined occurrences, and α is the sum over all α_i. The classification rule is then represented in Eq. (2).

$$l_{CNB}(d) = argmax\left[\log p(\overrightarrow{\theta_c}) - \sum_i f_i \log \frac{N_{\tilde{c}i} + \alpha_i}{N_{\tilde{c}} + \alpha}\right] \tag{2}$$

$\log p(\overrightarrow{\theta_c})$ is an assigned prior distribution over the set of classes, $f_i$is the count of occurrences of word *i* in document *d*. The negative sign indicates that documents poorly matching the classes other than *c* should be assigned to class *c* (Rennie et al. 2003).

Figure 1: Methodological procedure



Note: Models with different features are estimated at monthly frequency and news report frequency.

We estimate six different models, three using monthly aggregated features and three using high frequency data. In the aggregated analysis we use the monthly values for the macroeconomic indicators and month dummies and aggregate the news articles to one article per month. In the high frequency analysis, each news report is used as separate observation and each receives information on the macroeconomic indicators and month dummies from the current month. For both frequencies we use different combinations of feature classes. The first model each uses only the traditional information from months and macroeconomic variables as features, the second one uses only news articles and the third one uses all three feature classes. The methodological procedure is visualized in Figure 1.

In order to evaluate the different models, we calculate training and test set scores, which indicate the share of observations that are correctly classified. The results obtained from the best-performing model are graphically represented to allow for a better understanding of the model and we determine the most relevant features. With help of Shapley Additive Explanations (SHAP) plots (Lundberg and Lee 2017), we learn more about economic implications of our results, in particular

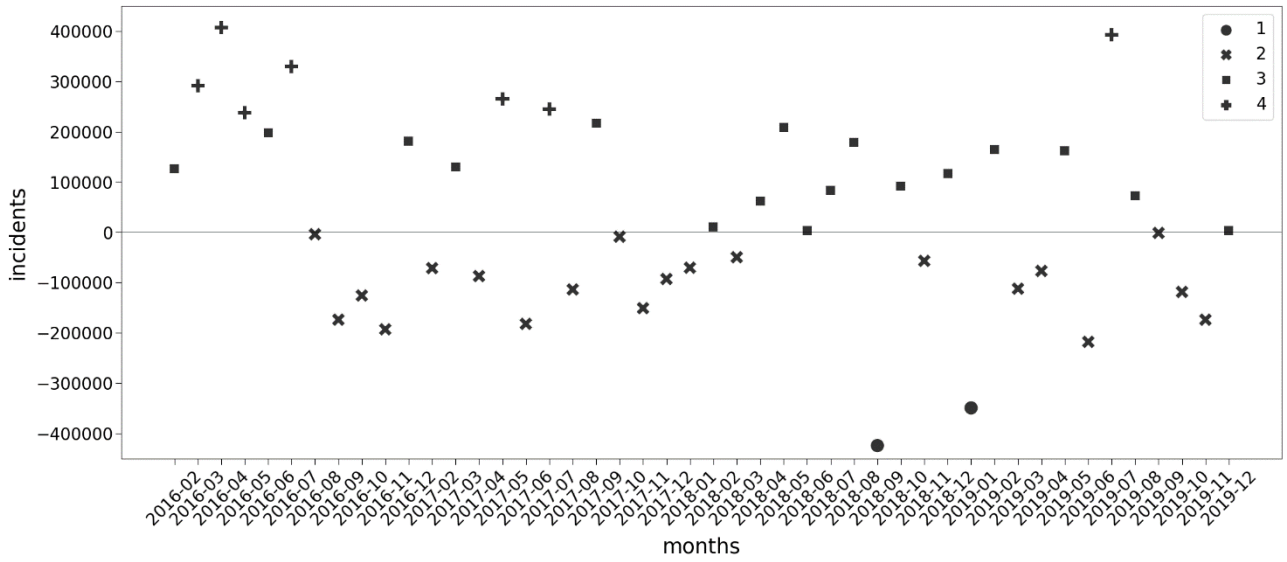
about correlations between features and target. As a robustness check we produce one-step-ahead forecasts for each specification by first using all available data until June 2019 and predicting the following month and then sliding the window one month forward several times until the data until November 2019 is used to predict December 2019.

6. Results

6.1. Descriptive results

In this section, the descriptive results are presented. The overall number of incidents per month according to the NCVS ranges between 1,250,000 and 2,850,000 in the period from the beginning of 2016 until the end of 2019. Figure 2 shows, for each month, the difference in the number of incidents compared to the previous month during that period. Due to differencing there is no observation for 2016-01. The observation for 2017-01 is categorized as an outlier and also excluded. For the remaining months, the difference in the number of incidents varies between -424,000, which constitutes the highest possible drop (percentage drop of 27 %), and 407,000, indicating the highest possible rise (percentage increase of 20 %). The markers show to which bin each observation belongs. Observations marked as circle fall into the first bin (-450,000 to -225,000), those marked as x fall into the second bin (-225,000 to 0), observations marked as square are part of the third bin (225,000 to 0) and those marked as plus are part of the fourth bin (225,000 to 450,000).

Figure 2: Development of differenced number of incidents from NCVS (target)



Note: Figure shows the difference in the number of incidents according to the NCVS compared to the previous month. No observation for 2016-01 (due to methodology) and 2017-01 (outlier). Markers indicate to which bin each observation belongs.

The following tables provide descriptive statistics for the aggregated analysis covering monthly observations and high frequency analysis covering observations at the frequency of the news

reports. The number of observations is lower for the target as explained above. After dropping the two respectively 1,546 observations corresponding to 2016-01 and 2017-01 for the aggregated respectively high frequency analysis, we have a balanced dataset consisting of 46 observations for the aggregated analysis and 35,428 observations for the high frequency analysis.

Table 1: Descriptive statistics target aggregated analysis

target	count	mean	stand. dev.	minimum	maximum
incidents differenced	46	28,761.46	187,236.76	- 424,107	406,671

Note: stand. dev. = standard deviation

The macroeconomic indicators represent changes compared to the previous month. The percentage change in GDP varies between -0.05 % and 0.09 %. The change in the consumer price index (CPI), varies between 0.83 and 2.95, the change in the unemployment rate lies between -0.3 and 0.2 during the considered period. Our month dummies take on either the value 0 or 1 and the mean is just 1/12.

Table 2: Descriptive statistics features aggregated analysis

class	features	count	mean	stand. dev.	minimum	maximum
macroeconomic indicators	GDP change in %	48	0.0142	0.0434	-0.0500	0.0900
	CPI	48	1.9120	0.5387	0.8271	2.9495
	unemployment change in %.	48	-0.0292	0.1166	-0.3000	0.2000
months	January	48	0.0833	0.2793	0	1
	February	48	0.0833	0.2793	0	1
	March	48	0.0833	0.2793	0	1
	April	48	0.0833	0.2793	0	1
	May	48	0.0833	0.2793	0	1
	June	48	0.0833	0.2793	0	1
	July	48	0.0833	0.2793	0	1
	August	48	0.0833	0.2793	0	1
	September	48	0.0833	0.2793	0	1
	October	48	0.0833	0.2793	0	1
	November	48	0.0833	0.2793	0	1
	December	48	0.0833	0.2793	0	1

class	features	count	unique	frequency
big data	news articles (aggregated)	48	48	1

Note: stand. dev. = standard deviation, GDP = normalized gross domestic product, CPI = consumer price index. GDP with less observations due to calculation of change compared to previous period.

In the high frequency analysis, the mean values are shifted due to the fact, that the number of news reports differs between the months. We can see that in January we have the lowest number of news reports and in March it reaches its maximum. Minimum and maximum values for the different features are as before.

Table 3: Descriptive statistics target high frequency analysis

target	count	mean	stand. dev.	minimum	maximum

incidents differenced	35,428	34,877.66	183,621.79	- 424,107	406,671

Note: stand. dev. = standard deviation

Table 4: Descriptive statistics features high frequency analysis

class	features	count	mean	stand. dev.	minimum	maximum
macroeconomic indicators	GDP change in %	36,974	0.0118	0.0429	-0.0500	0.0900
	CPI	36,974	1.8693	0.5463	0.8271	2.9495
	unemployment change in %.	36,974	-0.0282	0.1151	-0.3000	0.2000
months	January	36,974	0.0770	0.2666	0	1
	February	36,974	0.0764	0.2656	0	1
	March	36,974	0.0898	0.2859	0	1
	April	36,974	0.0802	0.2716	0	1
	May	36,974	0.0812	0.2732	0	1
	June	36,974	0.0865	0.2811	0	1
	July	36,974	0.0815	0.2736	0	1
	August	36,974	0.0846	0.2783	0	1
	September	36,974	0.0887	0.2844	0	1
	October	36,974	0.0981	0.2974	0	1
	November	36,974	0.0812	0.2732	0	1
	December	36,974	0.0748	0.2631	0	1

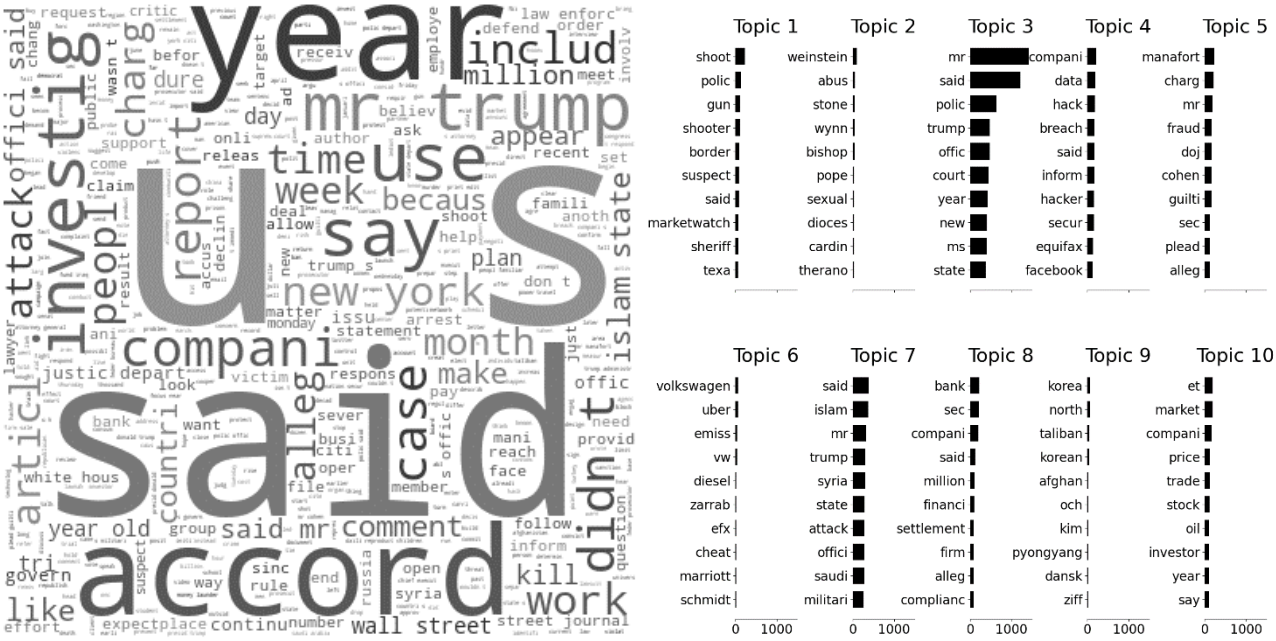
class	features	count	unique	frequency
big data	news articles	36,974	36,908	2

Note: stand. dev. = standard deviation, GDP = normalized gross domestic product, CPI = consumer price index. GDP with less observations due to calculation of change compared to previous period.

Looking more closely into the words from the DJN news articles, the 500 most frequent words are visualized in a word cloud in Figure 3A. *us*, *said, year* and *accord* are by far the most frequent ones. They are followed by *compani*, *say*, *use* and *mr trump*. Furthermore, there are several crime-related words among the most prevalent ones, e.g. *case*, *investig*, *charg* or *attack*. Figure 3B demonstrates the results of LDA with 10 different topics. The respective ten most frequent words of each topic are depicted in the diagram. While the first topic is mainly concerned with the police and shootings, the second one refers to publicly discussed cases of sexual abuse. Topic 3 is concerned with ex-president Trump, topic 4 with companies, data and hackers and topic 5 with Paul Manafort and fraud. Topic 6 refers to mobility and emissions, topic 7 to Islam and Muslim countries and 8 to banks and companies. Topic 9 is concerned with issues around (North) Korea, Taliban and Afghanistan and topic 10 with market developments and prices. Especially topic 3 and 7 consist of words that are frequently assigned to the topics.

Figure 3: Word cloud based on DJN articles

A *B*



Note: Words are pre-processed. 500 most frequent words used for word cloud (panel A). 10 most frequent words in 10 topics based on LDA are shown with the respective number of times each word was assigned to the respective topic.

6.2. Empirical results

The empirical results based on CNB Classifier demonstrate the significance of the text data for explaining the actual figure of crime (see Table 5). Our main indicator for ranking model performance is the test set score, which measures the extent to which the model correctly categorizes data that was not used to train the model (test data). Accuracy is then calculated by contrasting model predictions and actual outcomes for all observations in the test set and determining the share of correct predictions. At monthly frequency, the model using month dummies and monthly macroeconomic indicators achieves a test set score of almost 67 %, meaning that around two thirds of the data points can be classified correctly when the model categorizes previously unseen data. At this frequency, this value is above the score we achieve when using only the news articles (50 %) or combining all three feature classes (50 %). The benefit of the text data becomes apparent when estimating the model at the frequency of the news reports, using each report as a separate observation. In this specification, the model with only macro indicators and months achieves 63 % test set accuracy, the model using only news achieves 35 % and the model combining macro indicators, months and news delivers a test set accuracy of 70 %, which is the overall highest accuracy that could be achieved in this analysis. The high performance is confirmed, when comparing the accuracy of the model with the accuracy that could be achieved by always predicting the most frequent class (null accuracy). The null accuracy amounts to 49 %, being far below the accuracy of our best-performing CNB model. In addition, this approach also allows for the

construction of confidence intervals, e.g. when aggregating all predictions for one month to a mean value, possibly leading to improved estimates compared to the aggregated analysis.

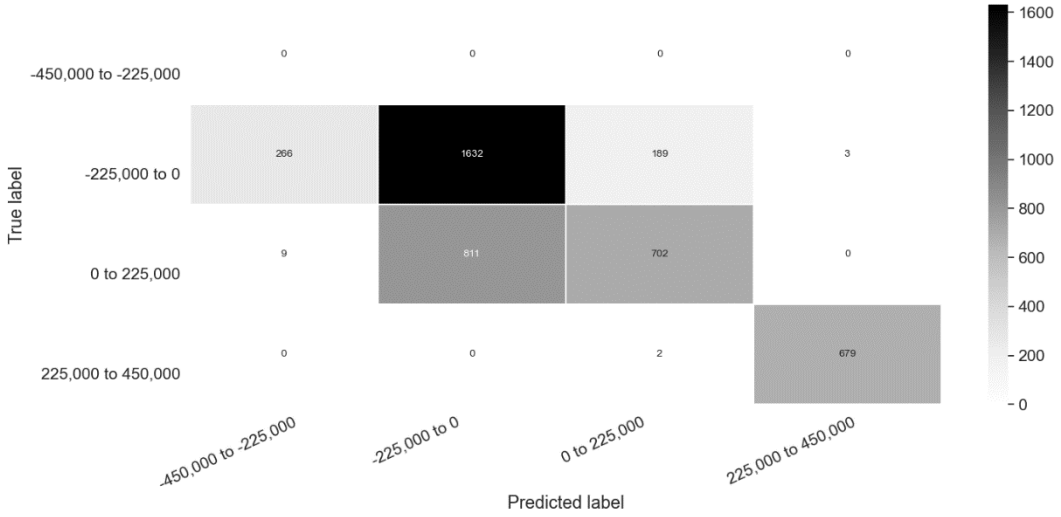
Table 5: Model specifications and results

Frequency	Monthly aggregates		High frequency	
Features / Accuracy	Training set	Test set	Training set	Test set
Macro indicators and months	0.7000	0.6667	0.6382	0.6308
News	0.4250	0.5000	0.6883	0.3499
Macro indicators and months + news	0.4750	0.5000	0.7818	0.7018

Note: Results from CNB Classifier.

In the following paragraphs we dive deeper into the best-performing model using all three data sources. Figure 4 depicts the distribution of predicted and true classes in a confusion matrix. The class -450,000 to -225,000 is not present in the test data and is predicted 266 times. The second class is predicted correctly 1,632 times, the third class 702 times with 820 predictions pointing at other classes while class 3 being true. The last class (225,000 to 450,000) is predicted correctly 679 times, with only 2 observations predicted as belonging to another class. Put into relation, the model struggles most with predicting whether observations belonging to the class 0 to 225,000 indeed belong to this class or to the class below, while it is quite precise in predicting the other classes correctly. Positively to note is that when the true class of an observation is a substantial increase in the true number of crimes, the model predicts this development with a very high accuracy. Since this category is the most alarming one, precisely predicting such developments is of particular interest.

Figure 4: Confusion matrix best-performing model



Note: Numbers in the matrix indicate frequencies with which the classes are predicted (in-)correctly in the test set.

In Table 6 we look at further performance metrics. Precision indicates the percentage of correct predictions of a specific class divided by all predictions of that class. Recall (also called sensitivity) measures the share of correct predictions of a specific class divided by all the occurrences of this

class in the data. The f1-score combines precision and recall into a harmonic mean by multiplying precision with recall times 2 and dividing by the sum of precision and recall. Support indicates the number of observations of each class in the training data. Since there are no observations for the first class, there are no precision metrics available for this class. The classification report mirrors the results from the confusion matrix. The model is particularly precise with respect to strong increases in the number of incidents, it performs well for predictions of slight reductions and has a high precision compared with lower recall for predictions of slight increases in the number of incidents.

Table 6: Classification report for best-performing model

Class	Precision	Recall	F1-score	Support
-450,000 to -225,000				0
-225,000 to 0	0.67	0.78	0.72	2090
0 to 225,000	0.79	0.46	0.58	1522
225,000 to 450,000	1.00	1.00	1.00	681

Note: No performance metrics for first class due to no observations from this class in the test data.

Furthermore, it is revealing to investigate predicted class probabilities (Figure 7 in the appendix). This allows for ranking the observations by the probability of it belonging to a specific class. In the CNB model, firstly the probabilities for the different classes are predicted and then the class with the highest predicted probability is chosen. Since there are four different classes, the threshold to predict one specific class is 0.25. The histograms for the first and fourth class are highly positively skewed with only few observations with a probability larger than 0.25. The histograms for the second and third class are more evenly distributed.

Compared to other ML methods, there are no classical hyperparameters to tune when using CNB. One possibility is to compare model results with different alpha values. Alpha is the additive smoothing parameter, indicating how to handle new words in the test data that have not appeared in the training data. Stepwise increasing alpha from 0 to 1 increases model performance considerably. Increasing alpha further continues to improve model performance at first, but reduces it when increasing alpha above 10. Since it is not recommended to raise alpha substantially above 1 since at some point the parameter rather than the occurrence of words starts to drive the probability of an observation belonging to a specific class, we decide to stick to the default of alpha equal to one. Applying a second normalization of the weights reduces model performance.

Since we used each news article separately, we made hundreds of predictions for the development of the number of incidents per month. In the next step, we aggregate all the individual predictions to aggregated forecasts for each month (Table 7) by counting the number of predictions of the respective bins in each month and selecting the most frequently predicted bin. The forecast is correct for 5 out of the 6 months and falls into the next lower category for August 2019.

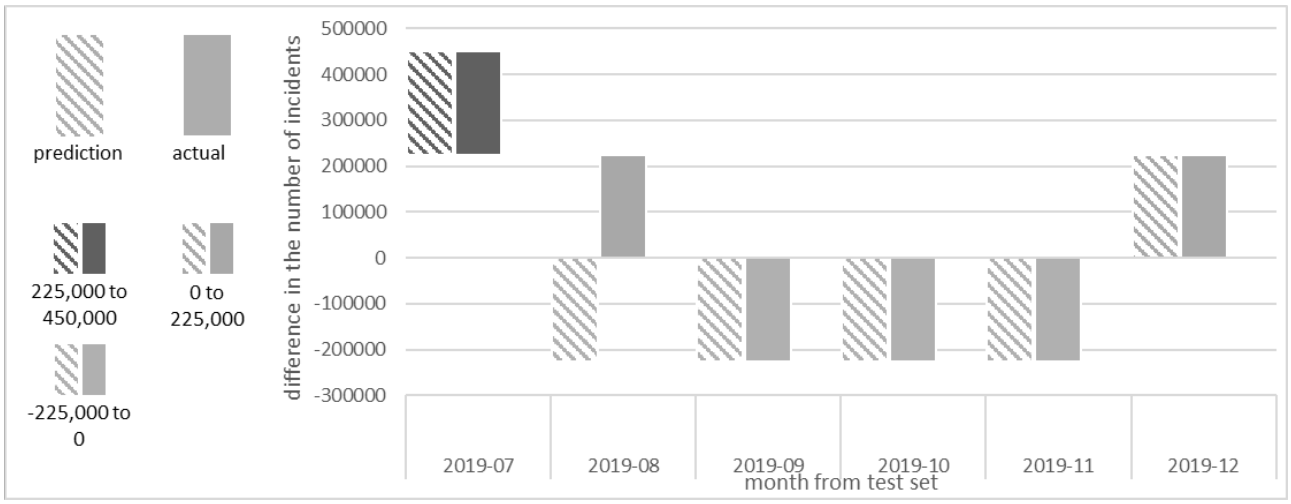
Table 7: Aggregated prediction results of best-performing model

month	count of bin predictions				predicted bin	actual bin
	bin 1	bin 2	bin 3	bin 4		
2019-07	0	0	2	**679**	4	4
2019-08	6	**801**	33	0	2	3
2019-09	266	**354**	125	0	2	2
2019-10	0	**646**	64	3	2	2
2019-11	0	**632**	0	0	2	2
2019-12	3	10	**669**	0	3	3

Note: Numbers in bold indicate bin with most predictions in the respective month.

Figure 5 visualizes the previous result and explains its meaning. In the first month of the test period (2019-07), the predicted bin for the development in the number of incidents is 225,000 to 450,000 (dark shaded bar). This corresponds to the actual bin (dark bar). In August 2019, the predicted development is a reduction between 0 and 225,000 (light shaded bar), whereas the actual development is an increase between 0 and 225,000 (light bar). In the following three months, the model correctly predicts a slight reduction in the number of incidents and in December 2019 it correctly predicts a slight increase.

Figure 5: Predictions and actual values in the test period



Note: Bars indicate intervals (bins) of the difference in the number of incidents in the respective month.

In order to check the robustness of our results, we performed one-step-ahead forecasts, meaning that we firstly use 3.5 years as training data (2016-02 until 2019-06) and predict the observations in the following month. We then move the window one step forward using the data until 2019-07 as training set and predicting August 2019. We continue with this sliding window until data up to November 2019 is used to train the model and December 2019 is predicted. Table 8 shows the results from this analysis. It becomes visible that the test scores vary between the different windows. Four out of the six months are predicted with extremely high accuracy, whereas the model is not able to correctly predict August 2019 and is in this specification quite unsure in September

2019. The average rounded prediction is correct for 4 out of 6 months and falls into the next lower category in the remaining two months. The mean test set score over all months is 72 % and thus slightly higher compared to the case where we used 3.5 years as training data for predictions of the following 6 months.

Table 8: One step ahead forecasts best-performing model

training period	test period	test score	count of bin predictions				predicted bin	actual bin
			bin 1	bin 2	bin 3	bin 4		
2016-02 - 2019-06	2019-07	0.9971	0	0	2	**679**	4	4
2016-02 - 2019-07	2019-08	0.0357	6	**712**	30	92	2	3
2016-02 - 2019-08	2019-09	0.3799	**311**	283	146	5	1	2
2016-02 - 2019-09	2019-10	0.9523	1	**679**	29	4	2	2
2016-02 - 2019-10	2019-11	1	0	**632**	0	0	2	2
2016-02 - 2019-11	2019-12	0.9663	3	20	**659**	0	3	3

Note: Numbers in bold indicate bin with most predictions in the respective month.

6.3. Validation of results

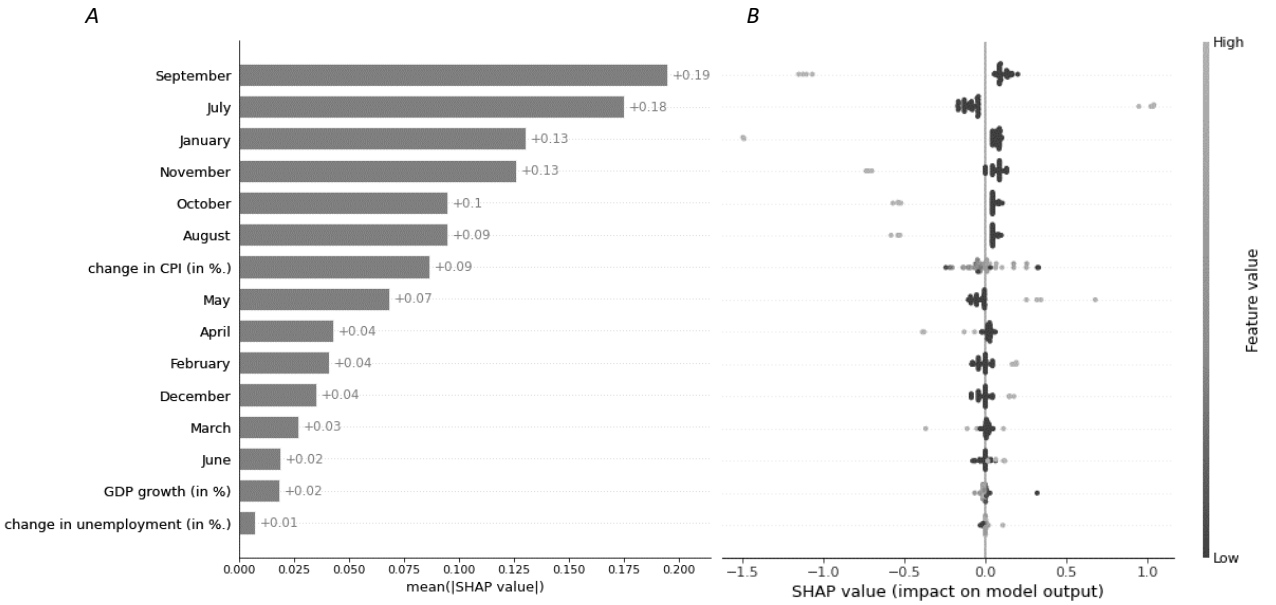
We expect our results to be stable as through the averaging over hundreds of predictions per month small errors in single predictions carry little weight. In our model, the most important factors are the month dummies and macroeconomic indicators. These two feature classes explain a substantial part of the variation. The addition of our text data helps to capture current developments from the media with which the predictions are refined. While we are able to light part of the dark through enabling earlier insights into criminal developments, some dark still remains. Firstly, there remains substantial uncertainty concerning the exhaustive number of incidents. As mentioned in previous chapters, the NCVS provides an additional perspective compared to official statistics that covers more incidents, however we cannot know with certainty how reliable this data is. On the one hand, it is not completely clear in how far survey errors might play a role and how honest victims respond to the question, whether they have been victimized. Still, since stating the truth is anonymous and does not bear direct consequences like having to testify or harming somebody, we expect the survey to capture substantially more incidents than recorded by the police. On the other hand, some crimes mentioned in the surveys might not be recognized as crimes by the police and, more importantly, there are various reasons why survey respondents might still not report all crimes to interviewers, so that we expect the disclosures in the surveys (and accordingly our predictions) to deviate from the true occurrences. Nevertheless, it is the best measure available. Secondly, it is not clear how the model would deal with profound and sudden changes in overall behaviour, e.g. during the COVID-19 pandemic. Finally, new dark might arise from possibly biased perceptions presented in the news articles under specific circumstances.

6.4. Economic and societal implications

To better understand which features drive the results, we looked at the 50 most important features for each category and found the macroeconomic indicators to be most relevant for all categories followed by the month dummies (Table 9 in the appendix). The most relevant words for each category align with some of the most frequent ones, namely *said*, *mr* and *trump*. They are followed by *polic*, *compani*, *offic(i)*, *investig* and *charg*. There are no substantial differences between the categories.

In order to better understand the direction of the impact of the macroeconomic and month variables, we compute SHAP values. We used the results from the aggregate model with only macroeconomic and month variables as features, since, due to the massive number of words that are used as features, incorporating the effects of this feature class was not possible. Panel A of Figure 6 displays the mean SHAP feature importances over all samples (training and test set) for the 15 features in a bar plot. In this specification, the month dummies obtain higher values compared to the macro variables. The highest importances are attributed to the features September and July, followed by January and November. The most important macroeconomic indicator is change in CPI, ranking 7th; GDP growth is the feature with the second least significance and change in unemployment is ranked last. Panel B summarizes the impact of the features on the model results in a beeswarm plot, which shows the impact higher and lower values of each feature have on the model output. The features are again ordered according to their importance and each dot in the row of a feature represents one observation. The scale on the right-hand side depicts the magnitude of the feature for this observation with light dots indicating high values and dark dots indicating low values.

Figure 6: SHAP bar and beeswarm plot



The results demonstrate that the development in the overall number of incidents exhibits a strong monthly pattern. In the month September (September = 1, which results in a high feature value and a light dot), we can expect fewer crimes (negative SHAP value), whereas in July, we expect higher crime rates, generally speaking. Looking at the macroeconomic indicators, increases in the CPI tend to increase crime rates. The effects of GDP and unemployment are ambiguous. There tends to be no substantial impact on model results from these two variables, with a slight tendency that lower GDP growth and higher growth in the unemployment rate increase crime rates. Although there is a highly significant negative correlation between GDP growth and developments in the unemployment rate, dropping one of the two variables does not increase test set accuracy. However, when estimating the model only with GDP and inflation, the positive (negative) correlation between GDP growth (changes in inflation) and the number of incidents becomes more apparent. Overall, there are thus indications of a negative correlation between crime rates and economic well-being.

6.5. Application to Europe and in particular Germany

Due to the broad availability of data with easy access, the US are frequently the centre of study when analysing unrecorded victimizations. Applying the methods for the US to Europe is rather difficult. First of all, data on victimization surveys is available to different extent in the different countries. The survey most comparable to the NCVS is the Crime Survey for England and Wales. Results from this survey can be obtained via the Office for National Statistics. Apart from that, there are the Swedish Crime Survey, where data is available for research purposes upon request from the Swedish National Council for Crime Prevention, Netherland´s Safety Monitor or the Scottish Crime

and Justice Survey, for which part of the data is made available. In general, when using the model to study different countries in parallel, factors that could influence truthful responses to the survey questions, such as trust in institutions, should be considered.

With the intention to apply part of the analysis also to Germany, we contacted the State Criminal Investigation Office (LKA) in Hamburg and learned that data on the dark figure of crime is gathered through victimization surveys as well. The first ones were conducted in 2012 and 2017 with the German Victimization Survey. Starting from 2020, it is planned to conduct a survey on victimization bi-annually, called Security and Crime in Germany. The data thereof, however, is not available for research purposes and only published as reports, thereby limiting the extent of possible insights substantially. We nevertheless looked at crime reporting rates in the country and found that they are quite similar compared to the US with rates between 32 % and 42 % (BKA 2022). In order to allow for in-depth research on crime patterns and especially unreported incidents, data access for researcher is required. Currently, scientists need to rely on countries with more established victimization surveys and less restrictive access for these types of analyses.

7. Discussion and conclusions

In this paper we explore a new approach for research on the actual figure of crime based on a machine learning model combining high-frequency information from crime-related news articles with more traditional information from macroeconomic indicators as well as monthly dummies. We provide an additional measure that allows to shed light into the dark of unregistered incidents up to two years earlier compared to surveys while being easy to use and relatively affordable. At the same time, our model is of high practical relevance to avoid economic costs arising from misallocations, to increase public safety and ultimately to increase prosperity.

Our analysis identifies text data from news reports as an additional source of valuable information for predictions of the actual extent of criminal activity in the US, that could be classified as some kind of aggregate of the perception of crime. With only this news source augmented by the three basic macroeconomic indicators GDP, unemployment and inflation as well as month dummies we are able to predict the development of criminal incidents as collected by the NCVS in the US with high accuracy of approximately 70 %. We detect that using non-aggregated news articles at their publication frequency is especially beneficial and that the best-performing model is a combination of more traditional features and novel high-frequency news data. Predicting only one month ahead slightly improves model performance on average with substantial differences between the months. Our results suggest that there is a strong monthly pattern present in the data and that our

macroeconomic variables add small pieces of information with a tendency towards lower crime levels with better economic performance.

By shedding light on the actual extent of criminal activity, including crimes that are not reported or detected, we provide benefits in several ways. (1) Efficient resource allocation in the police and increased willingness to allocate the necessary (financial) resources to police and protective measures. This could allow special focus groups during periods with a trend towards increases in crime, and prevent future crimes from occurring. (2) Early insights into crime patterns nationwide for decision makers as well as society. (3) Greater citizen confidence in statistics and thus in democracy. (4) Greater awareness of the issue in the public debate, which could encourage citizens to report crimes.

The results from our analysis allow for direct conclusions concerning the dark figure of crime, since the data on monthly incidents consists mainly of unreported cases and increases in the overall number of incidents are related to increases in the number of non-reported incidents. Thus, reducing overall crime rates as reported in the NCVS should reduce the number of unreported crimes as well. Exact effect mechanisms of particular intervention measures on the dark figure of crime should be the centre of future research to allow for targeted actions.

There are several suggestions for further development of the model. An analysis combining different news data or even other big data sources could be revealing. In addition, further applications could allow for segmentation into different states or cities and highlight more regional features or for segmentation into different types of crime, e.g. depending on reporting probabilities or caused harm. This paper shows that it is possible to get closer to the actual figure of crime using the tools presented and it is possible to adjust the model to the specific needs of its users.

Publication bibliography

Babuta, Alexander (2017): Big Data and Policing. An Assessment of Law Enforcement Requirements, Expectations and Priorities. Edited by Royal United Services Institute for Defence and Security Studies (RUSI). Available online at https://static.rusi.org/201709_rusi_big_data_and_policing_babuta_web.pdf.

Becker, Gary S. (1968): Crime and Punishment: An Economic Approach: The University of Chicago Press (76). In *Journal of Political Economy* (2), pp. 169–217.

Biderman, Albert D.; Reiss, Albert J. (1967): On Exploring the "Dark Figure" of Crime (374). In *The ANNALS of the American Academy of Political and Social Science* (1), pp. 1–15.

Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. (2003): Latent Dirichlet Allocation (3). In *Journal of Machine Learning Research* (Jan), pp. 993–1022.

Buil-Gil, David; Medina, Juanjo; Shlomo, Natalie (2021): Measuring the Dark Figure of Crime in Geographic Areas: Small Area Estimation from the Crime Survey for England and Wales (61). In *British Journal of Criminology* (2), pp. 364–388.

Castelbajac, Matthieu de (2014): Brooding Over the Dark Figure of Crime (54). In *British Journal of Criminology* (5), pp. 928–945.

Chalfin, Aaron; McCrary, Justin (2017): Criminal Deterrence: A Review of the Literature (55). In *Journal of Economic Literature* (1), pp. 5–48.

Coleman, Clive; Moynihan, Jenny (1996): Understanding Crime Data. Haunted by the Dark Figure. Buckingham: Open University Press (Crime and justice).

Curiel, Rafael P.; Cresci, Stefano; Muntean, Cristina I.; Bishop, Steven R. (2020): Crime and its Fear in Social Media (6). In *Palgrave Communications* (1).

Draca, Mirko; Machin, Stephen (2015): Crime and Economic Incentives (7). In *Annual Review of Economics* (1), pp. 389–408.

Ehrlich, Isaac (1973): Participation in Illegitimate Activities: A Theoretical and Empirical Investigation: The University of Chicago Press (81). In *Journal of Political Economy* (3), pp. 521–565.

Estrada, Felipe (2006): Trends in Violence in Scandinavia According to Different Indicators. An Exemplification of the Value of Swedish Hospital Data (46). In *British Journal of Criminology* (3), pp. 486–504.

Federal Criminal Police Office (BKA) (2022): Kriminalstatistisch-kriminologische Analysen und Dunkelfeldforschung. Available online at https://www.bka.de/DE/UnsereAufgaben/Forschung/ForschungsprojekteUndErgebnisse/Dunkelfeldforschung/dunkelfeldforschung_node.html.

Freeman, Richard B. (1999): The Economics of Crime. In Orley C. Ashenfelter, David Card (Eds.): Handbook of Labor Economics. Amsterdam, Boston, Heidelberg, London, New York, Oxford, Paris, San Diego, San Francisco, Singapore, Sydney, Tokyo: North-Holland (Handbooks in economics, 3C), pp. 3529–3571. Available online at https://www.sciencedirect.com/science/article/pii/s1573446399300432.

Gerber, Matthew S. (2014): Predicting Crime using Twitter and Kernel Density Estimation (61). In *Decision Support Systems*, pp. 115–125.

Goudriaan, Heike; Wittebrood, Karin; Nieuwbeerta, Paul (2006): Neighbourhood Characteristics and Reporting Crime. Effects of Social Cohesion, Confidence in Police Effectiveness and Socio-Economic Disadvantage (46). In *British Journal of Criminology* (4), pp. 719–742.

Kaufmann, Mareile; Egbert, Simon; Leese, Matthias (2019): Predictive Policing and the Politics of Patterns (59). In *British Journal of Criminology* (3), pp. 674–692.

Kennedy, Leslie W. (1988): Going it Alone: Unreported Crime and Individual Self-help (16). In *Journal of Criminal Justice*, 1/1/1988 (5), pp. 403–412. Available online at https://www.sciencedirect.com/science/article/pii/0047235288900657.

Lundberg, Scott M.; Lee, Su-In (2017): A Unified Approach to interpreting Model Predictions (30). In *Advances in Neural Information Processing Systems*, pp. 4765–4774.

MacDonald, Ziggy (2001): Revisiting the Dark Figure: A Microeconometric Analysis of the Underreporting of Property Crime and its Implications (41). In *British Journal of Criminology* (1), pp. 127–149.

National Archive of Criminal Justice Data (NACJD) (2022): National Crime Victimization Survey. University of Michigan. Available online at https://www.icpsr.umich.edu/web/pages/NACJD/NCVS/index.html, checked on 12.04.22, 11:31.

Oatley, Giles C. (2022): Themes in Data Mining, Big Data, and Crime Analytics (12). In *WIREs Data Mining and Knowledge Discovery* (2).

Organisation for Economic Co-Operation and Development (OECD) (2022a): Inflation (CPI). Available online at https://data.oecd.org/price/inflation-cpi.htm, updated on 29.08.22, 14:16.

Organisation for Economic Co-Operation and Development (OECD) (2022b): Monthly Economic Indicators. Composite Leading Indicators (MEI). Normalised (GDP). Available online at https://stats.oecd.org/, checked on 29.08.22, 11:32.

Porter, Lauren C.; Curtis, Andrew; Jefferis, Eric; Mitchell, Susanne (2020): Where's the Crime? Exploring Divergences between Call Data and Perceptions of Local Crime (60). In *British Journal of Criminology*, pp. 444–467.

Rennie, Jason D.; Shih, Lawrence; Teevan, Jaime; Karger, David R. (2003): Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 616–623. Available online at https://www.aaai.org/papers/icml/2003/icml03-081.pdf?ref=https://githubhelp.com.

Shah, Neil; Bhagat, Nandish; Shah, Manan (2021): Crime Forecasting: a Machine Learning and Computer Vision Approach to Crime Prediction and Prevention (4). In *Visual Computing for Industry, Biomedicine, and Art* (9), pp. 1–14.

Skogan, Wesley G. (1977): Dimensions of the Dark Figure of Unreported Crime (23). In *Crime & Delinquency* (1), pp. 41–50.

Skogan, Wesley G. (1984): Reporting Crimes to the Police: The Status of World Research (21). In *Journal of Research in Crime and Delinquency* (2), pp. 113–137.

Smith, Gavin J.; Bennett Moses, Lyria; Chan, Janet (2017): The Challenges of doing Criminology in the Big Data Era: Towards a Digital and Data-driven Approach (57). In *British Journal of Criminology* (2), pp. 259–274.

U.S. Bureau of Labor Statistics (U.S. BLS) (2022): Labor Force Statistics from the Current Population Survey. (Seas) Unemployment Rate. Available online at https://beta.bls.gov/dataViewer/view/timeseries/LNS14000000.

United States. Bureau of Justice Statistics (U.S. BJS) (2020a): National Crime Victimization Survey, [United States], 2016. Edited by Inter-university Consortium for Political and Social Research. DOI: 10.3886/ICPSR36828.v4.

United States. Bureau of Justice Statistics (U.S. BJS) (2020b): National Crime Victimization Survey, [United States], 2017. Edited by Inter-university Consortium for Political and Social Research. DOI: 10.3886/ICPSR36981.v2.

United States. Bureau of Justice Statistics (U.S. BJS) (2020c): National Crime Victimization Survey, [United States], 2018. Edited by Inter-university Consortium for Political and Social Research. DOI: 10.3886/ICPSR37297.v1.

United States. Bureau of Justice Statistics (U.S. BJS) (2020d): National Crime Victimization Survey, [United States], 2019. Edited by Inter-university Consortium for Political and Social Research. DOI: 10.3886/ICPSR37645.V1.

United States. Bureau of Justice Statistics (U.S. BJS) (2021a): National Crime Victimization Survey, [United States], 2020. Edited by Inter-university Consortium for Political and Social Research. DOI: 10.3886/ICPSR38090.v1.

United States. Bureau of Justice Statistics (U.S. BJS) (2021b): National Crime Victimization Survey, [United States], 2020. Codebook. Edited by Inter-university Consortium for Political and Social Research.

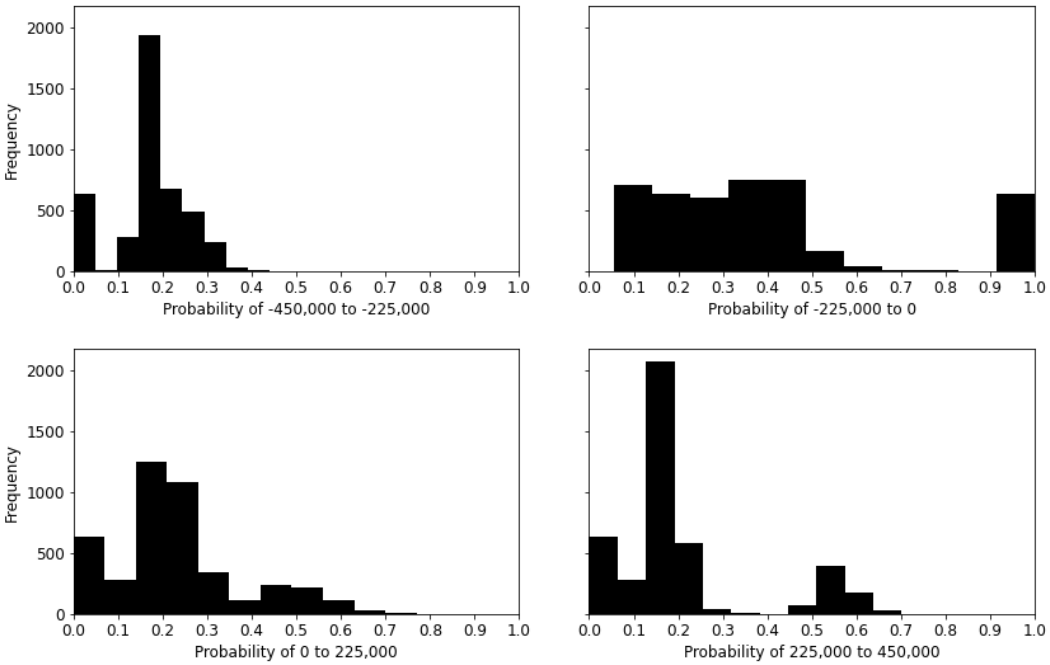
Wang, Hongjian; Kifer, Daniel; Graif, Corina; Li, Zhenhui (2016): Crime Rate Inference with Big Data. In Balaji Krishnapuram, Mohak Shah, Alex Smola, Charu Aggarwal, Dou Shen, Rajeev Rastogi (Eds.): Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA, 13 08 2016 17 08 2016. New York, NY, USA: ACM, pp. 635–644.

Williams, Matthew L.; Burnap, Pete; Sloan, Luke (2017): Crime Sensing with Big Data: The Affordances and Limitations of using Open Source Communications to estimate Crime Patterns (57). In *British Journal of Criminology*, 320-340.

Zhang, Harry (2004): The Optimality of Naive Bayes (1). In *Aa* (2).

Appendix

Figure 7: Histograms for probabilities of predicting a specific class



Note: Histograms indicate frequencies for the probabilities of predicting a specific class based on the CNB classifier.

Table 9: Feature importances for the different categories

bin 1 -450,000 to -225,000	bin 2 -225,000 to 0	bin 3 0 to 225,000	bin 4 225,000 to 450,000
change in unempl.	change in unempl.	change in unempl.	change in CPI
change in CPI	change in CPI	change in CPI	change in unempl.
GDP growth	GDP growth	GDP growth	GDP growth
March	May	March	June
June	July	November	October
May	February	April	February
April	March	October	September
October	June	July	March
February	December	September	November
November	September	May	April
July	April	August	August
August	said	June	December
December	mr	January	mr
September	October	said	said
said	August	mr	May
mr	January	December	January
trump	trump	February	trump
January	polic	polic	compani
polic	state	trump	year
year	compani	year	July
state	year	state	polic
compani	offic	compani	state
offici	offici	offici	offici

offic	charg	offic	say
say	investig	new	charg
new	say	say	new
charg	new	court	investig
investig	court	investig	court
court	alleg	attack	offic
report	report	charg	report
alleg	bank	report	alleg
attack	attack	alleg	case
case	case	case	presid
peopl	presid	peopl	peopl
bank	govern	depart	attack
presid	secur	govern	secur
depart	peopl	secur	govern
govern	depart	bank	depart
secur	feder	citi	bank
feder	justic	presid	feder
law	law	feder	ms
ms	prosecutor	law	law
justic	judg	accord	million
million	attorney	islam	accord
accord	million	million	justic
prosecutor	ms	ms	prosecutor
shoot	accord	kill	attorney
attorney	use	prosecutor	use
citi	shoot	justic	judg
use	sec	use	includ