Can Bad News Be Good Predictors?

Estimating Actual Crime Levels with Crime-Related News

Christina H. Maass
University of Hamburg

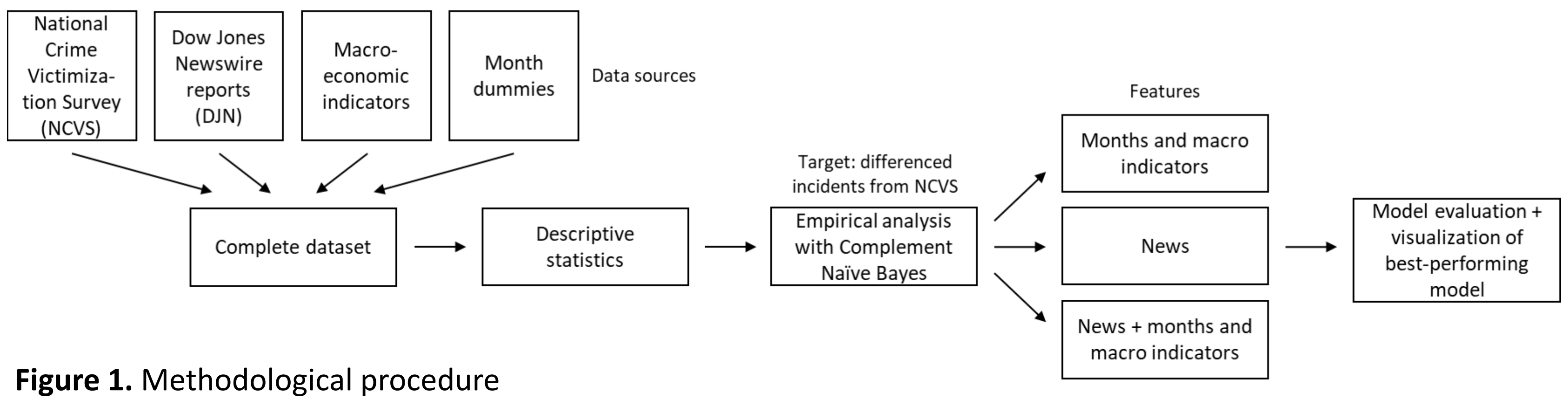
Abstract

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 of America (**US**) as indicated by the **National Crime Victimization Survey (NCVS)** timely and with high accuracy. Our approach enables improvements in resource allocation, increased public safety and thus greater economic prosperity.

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**. The **main factor** seems to be simple **non-reporting by citizens** as already found by [1] and confirmed by low reporting rates in recent victimization surveys. Therefore, unreported crimes are the focus of this paper.

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.



National Crime Victimiza-tion Survey (NCVS)
Dow Jones Newswire reports (DJN)
Macro-economic indicators
Month dummies
Data sources
Complete dataset
Descriptive statistics
Target: differenced incidents from NCVS
Empirical analysis with Complement Naïve Bayes
Features
Months and macro indicators
News
News + months and macro indicators
Model evaluation + visualization of best-performing model

Figure 1. Methodological procedure

Data and Method

Approximation for the actual figure of crime: total number of incidents (including unreported ones) from the **NCVS** in the US in the years **2016-2019**

Target: difference between incidents in current and previous month binned into 4 bins [-450,000 to -225,000], [-225,000 to 0], [0 to 225,000], [225,000 to 450,000]

Features:

(1) Monthly values for the macro indicators GDP, unemployment and inflation
(2) Dummies for each month of the year
(3) Dow Jones Newswire (DJN) reports (economic & financial news from different news services published by Dow Jones & Company)

To extract information from the highly frequent and unstructured news data we use a **machine learning** approach based on **Complement Naïve Bayes (CNB)** [2] due to its high performance when using text data. We estimate 6 different models (3 aggregated and 3 high frequency) with different combinations of feature classes. In the **aggregated analysis** we use the monthly values for the macro and month variables 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 macro and month variables from the current month. The **training period** is **01-2016 – 06-2019** and the **test period** is **07-2019 – 12-2019**.

Our main indicator for ranking model performance is the **test set score** (share of correctly categorized previously unseen data points). In order to better understand the direction of the impact of the macroeconomic and month variables, we compute **Shapley Additive Explanations (SHAP) values** [3].

Contact

Christina H. Maass, University of Hamburg
✉ christina.maass@uni-hamburg.de

Website
uhh.de/wiso-maass-en

Full paper
hdl.handle.net/10419/277607

Results

The **benefit of the text data** becomes apparent when estimating the model with the CNB Classifier at the **(high) frequency of the news reports** (Table 1). 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 all three sources** delivers the overall highest test set **accuracy of 70 %.**

In the next step, we aggregate all individual predictions per month to aggregated forecasts for each month by selecting the most frequently predicted bin each. The **forecast is correct for 5 out of the 6 months** and falls into the next lower category for August 2019 (Figure 2).

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

Table 1. Results with CNB classifier in different specifications.

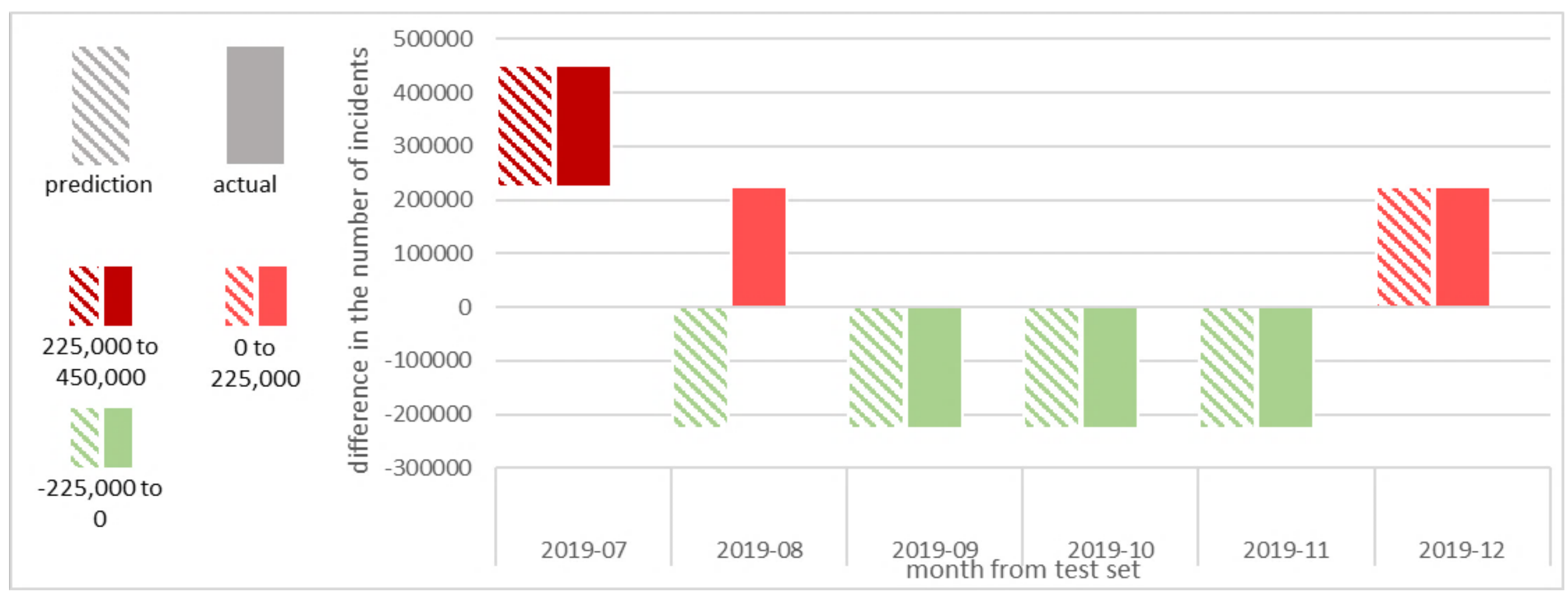


Figure 2. Predictions and actual bins in the test period.

Panel A of Figure 3 displays the mean **SHAP feature importances** over all samples (training and test set) for the 15 features in a bar plot. The **month dummies obtain higher values compared to the macro variables**. Panel B summarizes the impact higher and lower values of each feature have on the model output in a **beeswarm plot**. Each dot in the row of a feature represents one observation.

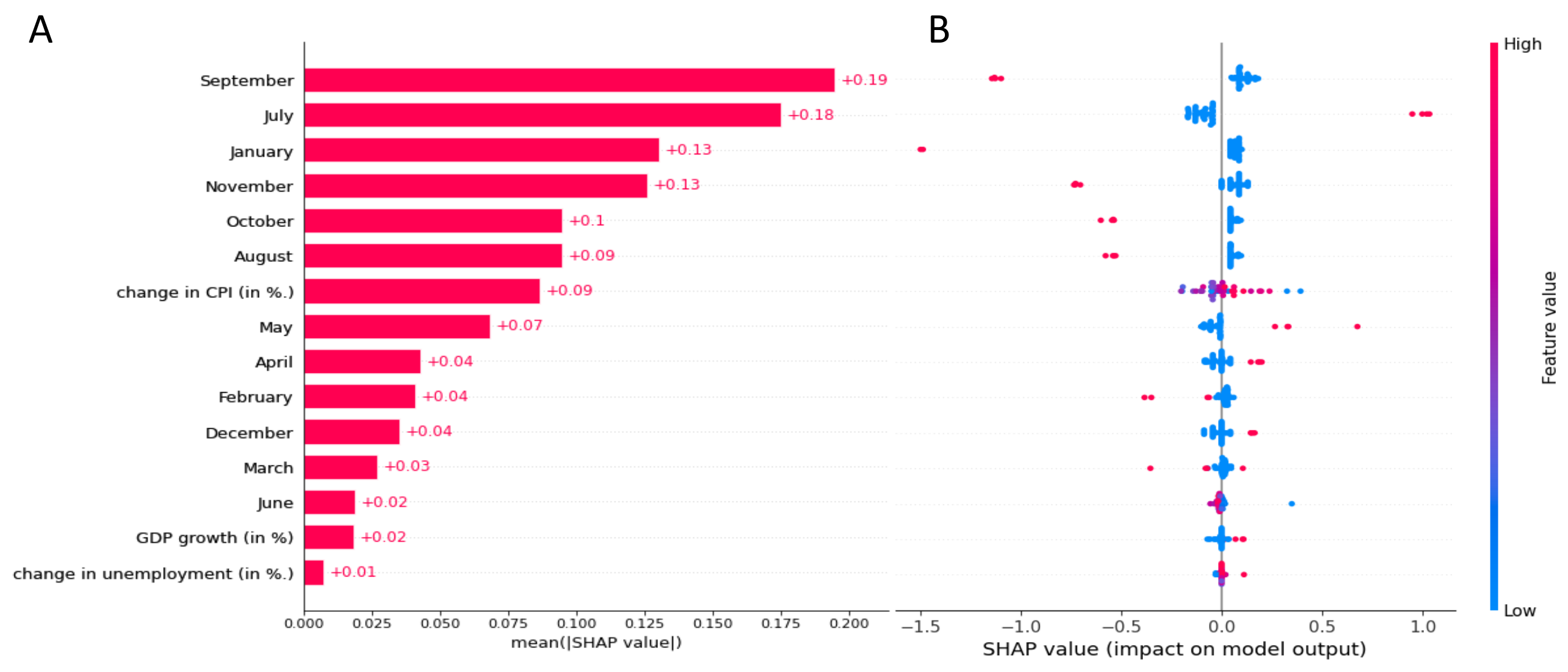


Figure 3. SHAP bar and beeswarm plot.

Discussion

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. 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. Secondly, it is not clear how the model would deal with profound and sudden changes in overall behavior, e.g. during the COVID-19 pandemic. Finally, new dark might arise from possibly biased perceptions presented in the news articles under specific circumstances.

Conclusions

By shedding light on the actual number of crimes, we provide several benefits:

(1) Efficient resource allocation in the police and increased willingness to allocate the necessary (financial) resources to police and protective measures
(2) Early insights into crime patterns nationwide for decision makers and 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

References

1. Skogan, Wesley G. (1977): Dimensions of the Dark Figure of Unreported Crime (23). In Crime & Delinquency (1), pp. 41–50.
2. 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.
3. 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.