# Demanding justice in the cloud. An analysis of punitive attitudes in social media for traditional and cyber-enabled crime.

This report accounts for the results obtained in the project "Demanding justice in the cloud. An analysis of punitive attitudes in social media for traditional and cyber-enabled crime", funded by the Centre for Digital Trust and Society at the University of Manchester through its seed corn funding. The project aims to harness computational criminology tools (Ozalp et al. 2020; Williams et al. 2020) to analyse online punitive attitudes between 2019 and 2022 in the UK. The team, composed of criminology and computer science department members, focused on data collection and using various Natural Language Processing (NLP) techniques to detect punitive communications on Twitter automatically.

## Introduction

The study of punitive attitudes is an area of growing interest in criminology and the social sciences (Adriaenssen and Aertsen 2015; Aguilar-Jurado 2018). Punitive attitudes involve the support of harsher penal punishments, and it has been shown that they can significantly impact the functioning of the criminal justice system and crime prevention policies (Enns 2014; Jennings et al. 2017). Instead of following expert advice and empirical evidence, populist politicians are more guided by the opinions and emotional reactions of the public (Bottoms 1995; Garland 2021). Some of these policies that involve higher use of imprisonment and longer sentences not only do not work to prevent or reduce crime but might involve several negative externalities in terms of recidivism, human rights and economic costs (Cullen et al. 2011; Liu et al. 2018; Loeffler and Nagin 2022).

In the UK, even more than 50 years after the death penalty abolition, it is still supported by 40% of the population (Kirk 2022a). When asked about some specific crimes (murder of a child, acts of terrorism, or multiple murders), the death penalty is supported by a majority of the population (ibid). In addition, approximately two-thirds of Britons believe that the sentences received by offenders convicted of crimes are not tough enough (Kirk 2022b). This public opinion regarding punishment takes place in a context of growing economic uncertainty (KPMG 2022) and an increase in violence (pre-pandemic) that had been steadily declining since the mid-1990s (Office for National Statistics, 2021; Office for National Statistics, 2022). Furthermore, the recent covid-19 pandemic and its lockdown affected social interactions in general and criminal strategies in particular, multiplying cyber-dependent and cyber-enabled crime (Buil-Gil et al. 2021), raising the question of the effects of this type of crime on citizens' attitudes.

All these elements raise the question of the transformations that took place regarding the citizens' punitive attitudes during this period. Traditionally, research on punitive attitudes has been carried out with data from public opinion surveys. This type of data has the advantage of statistical representativeness, allowing researchers to infer results from the sample to the larger population. However, such surveys usually involve a static snapshot that is not sensitive to the dynamics of changing and real-time flows of citizens' opinions over weeks and months (Williams et al., 2020). Another important aspect is that punitiveness surveys are focused on atomised individuals and hardly manage to capture existing networks of communication, exchange, and discussion. Finally, the use of surveys to measure sensitive or controversial issues may suffer from social desirability bias where the respondent tailors their response to what they feel is expected by the interviewer or the community (Stephens-Davidowitz 2018).

This project takes advantage of the naturally occurring communications on social media to generate a measure of citizens' punitive attitudes online. The use of this type of data is an important complement to traditional data, filling the aforementioned gaps that the survey data implies, making it possible to generate low-cost longitudinal data to explore this topic. While not representative of the population, these data can be combined with other data sources to generate a more complete picture of the phenomenon. Moreover, in a global context where a significant part of public discussion is generated on social media, understanding how punitive attitudes are developed and disseminated online is also intrinsically relevant (Gerbaudo 2014; Bail et al. 2017; Hartman et al. 2022).

## Punitive Attitudes

The concept of punitive attitudes remains ‘under-theorised’ and vague (Adriaenssen and Aertsen 2015; Aguilar-Jurado 2018), and there are still strong disagreements in the field regarding what it means and how it should be measured. Most studies use the term ‘attitudes’, but only Maguire and Johnson (2015) explicitly discuss the concept in the context of cognitive and social psychology. According to this literature, punitive attitudes can be considered as one specific type of evaluative attitude (Ajzen and Fishbein 2005; Riemer et al. 2014; Albarracin and Shavitt 2018), that is, an individual tendency to have a positive or approval reaction to a specific psychological object: criminal justice policies or institutions that promote the increase of the costs of crime for offenders through the application of punishments (Aguilar-Jurado 2018; Brooks 2021).

The development of studies on punitive attitudes has advanced significantly since its origins (Wood and Viki 2001; Aguilar-Jurado 2018). These studies originally included only very general measures of attitudes to sentencing and punishment, usually based on a single item and with very little context (e.g. through support for the death penalty or a general increase in sentences). Such studies may overestimate the punitive attitudes of individuals, as they do not provide information on other types of measures (such as rehabilitative measures) and do not consider the characteristics of the crime, the victim, or the offender (Stalans 2013).

The very development of research led to the sophistication of measurement strategies. First, several studies included specific questions related to concrete cases using vignettes (case-scenario). These studies allow not only to control for crime characteristics but also to manipulate them in experimental contexts (Campregher and Jeglic 2016; Jahnke 2018; Socia et al. 2019; Horstman et al. 2021). Other studies have used batteries of questions to strengthen the validity of the measures used through psychometric analysis strategies (Ortet-Fabregat and Pérez 1992; Mascini and Houtman 2006; Piquero et al. 2010; Ramirez 2013; Aizpurúa 2015; Maguire and Johnson 2015; Armborst 2017). Finally, qualitative strategies have also been explored, albeit to a lesser extent, for analysing punitive attitudes (Cook and Powell 2003; Boda and Szabó 2011; Leverentz 2011).

However, methodological sophistication is not without its limitations. For example, surveys and interviews usually involve a static snapshot that is not sensitive to the dynamics of changing and real-time flows of citizens' opinions over weeks and months (Williams et al. 2020). Another important aspect is that punitiveness surveys are focused on atomised individuals and hardly manage to capture existing networks of communication, exchange, and discussion. Additionally, using surveys to measure sensitive or controversial issues may suffer from social desirability bias, where the respondent tailors their response to what they feel is expected by the interviewer or the community (Stephens-Davidowitz 2018). Finally, the use of social media quite effectively captures the behavioural dimension of attitudes insofar as the decision to engage in this type of communication arises from the interaction between the user and the network, without the researcher's interference.

## Computational Social Science

As previously suggested, some of the limitations of surveys can be addressed with the complementary use of Emerging and New Forms of Data (NEFD). NEFDs are multimedia data recorded digitally through different types of connections (Radanliev and De Roure 2023). The expansion of the Internet for personal use and its application in various kinds of processes and objects led to an exponential growth of the data available within this category (Bail 2014; Edelmann et al. 2020). NEFDs offer detailed information about the development of social relationships in large populations as they occur, unlike the traditional datasets that social scientists typically collect (Bail 2014; Edelmann et al. 2020; Hofman et al. 2021; Radanliev and De Roure 2023).

At the same time, both the computational power and the techniques for analysing these data streams developed in parallel with the multiplication of NEFDs. This intersection between new analysis techniques and data gave rise to the new interdisciplinary field called Computational Social Science (Edelmann et al. 2020).

Particularly relevant to this project is the background in the use of social media for the analysis of collective behaviour and social psychology. Some of the most pertinent antecedents to the project are presented below.

### Twitter Data for the study of Collective Behaviour and Social Attitudes

Computational social science has motivated significant research on collective behaviour and politics. Computational criminology has significantly advanced in this field (Williams and Burnap 2016), mainly through studies on hate and cyber-hate (Ozalp et al. 2020; Williams et al. 2020). These studies have in common the use of textual data from the social network Twitter to capture hate messages. In addition, these projects take advantage of NEFD from social media to account for the fluctuation in online hate and its causes and consequences. These studies also pioneered techniques traditionally associated with Data Science for estimating the measure of interest through the use of Artificial Intelligence for automatically classifying tweets.

Another interesting element is the connection between online and offline events. The analysis of the causal direction has been carried out in both directions: accounting for the effects of online attitudes on the offline world and of offline events on the online world. Thus Williams (2020) uses geo-referenced hate tweets to predict offline hate crime in London, and conversely, Williams and Burnap (2016) analyse the online effect of the terrorist acts in Woolwich 2013 in cyber-hate against Black Minority Ethnic (BME) and religious groups. Powell et al (2018) analysed responses to crime through the case study of the disappearance, abuse, and murder of Jill Meagher on Twitter. Kostakos (2018) investigated the discussion of organised crime on social media using data from Google and Twitter. Finally, (O’Connor 2017) analysed the type of communications carried out by police departments in Canada.

However, such studies have not been limited to criminology. Flores (2017) utilises textual data from Twitter to demonstrate how the enactment of anti-immigration laws in Arizona led to a hardening of public attitudes towards immigration, with a quasi-experimental design. Several studies used online data to track various elements of interest during the Covid-19 pandemic worldwide. From information flow mapping (Prabhakar Kaila and Prasad 2020) to state leaders' responses (Haman 2020; Rufai and Bunce 2020), including disinformation (Singh et al. 2020; Yang et al. 2020) and emotional reactions to it (Xue et al. 2020).

This body of literature offers clues as to how it is possible to use data from social media to complement other data sources.

## Methods

### Data Collection

All the information used in this study was collected from the online social media platform Twitter. Twitter allows users to share short messages known as "tweets" of up to 280 characters. Launched in 2006, it has become a popular means of communication for individuals and organisations, with over 330 million monthly active users in 2017. Its format emphasises brevity and real-time communication, encouraging users to share thoughts, ideas, and updates with their followers. In addition, users can follow other accounts to see their tweets in their timeline and use hashtags and mentions to participate in broader conversations on specific topics or with other users (Leis et al. 2019; Britannica 2023).

The ease of access to its data for researchers, as well as its relevance to news and political communication, has led to a proliferation of social science studies using the platform (Jin et al. 2014; Himelboim et al. 2016; Flores 2017; Leis et al. 2019; Ozalp et al. 2020; Prieto Curiel et al. 2020; Williams et al. 2020; Hu and Kearney 2021).

The data was downloaded via Twitter's academic API between February and March of 2023. The connection was made through the statistical software R and the package `academictwitteR` (Barrie and Ho 2021).

In order to analyse reactions to crime in the United Kingdom, the following strategy was carried out. First, a selection of the most important British press outlets on Twitter was made. For these media, all tweets related to crime news were downloaded based on a selection of keywords. Then, all user responses to these news items were downloaded. These responses were manually scored as punitive or non-punitive based on a set of rules that were defined by the research team. This set of annotated tweets was used to train a large set of Machine Learning models and Transformers for the automatic detection of punitive tweets. Once this experimentation had been carried out and the model with the best performance was selected, the prediction was made on the total base of responses, obtaining the time series of punitive reactions to the crime (Figure 1).

Figure 1 Data collection and classification pipeline



For the news download, a sample of 20 news portals was selected.

Table 1 News Sources

|  |  |  |
| --- | --- | --- |
| **News Source** | **User** | **Followers** |
| BBC News (UK) | BBCNews | 15090513 |
| The Guardian | guardian | 10895497 |
| Sky News | SkyNews | 8386716 |
| The Independent | Independent | 3628283 |
| The Telegraph | Telegraph | 3328749 |
| Daily Mail Online | MailOnline | 2821249 |
| The Sun  | TheSun | 2065979 |
| The Times and The Sunday Times | thetimes | 1771013 |
| The Mirror | DailyMirror | 1386788 |
| BBC Scotland News | BBCScotlandNews | 607146 |
| Liverpool Echo | LivEchonews | 534157 |
| Birmingham Live | birmingham\_live | 351991 |
| Wales Online | WalesOnline | 345981 |
| BBC News NI | BBCNewsNI | 336178 |
| Daily Star | dailystar | 234632 |
| BBC South East | bbcsoutheast | 230662 |
| Evening Standard | EveningStandard | 165954 |
| BBC East Yorks and Lincs | looknorthBBC | 87160 |

All tweets with at least one of the following queries (grouped by crime) between 2019 and 2022 were downloaded for each account. Only tweets mentioning any crime or proxy (e.g., illegal substances) were selected. The selection of words and crime types was based on an initial exploratory analysis using Twitter, LexisNexis, and Media Cloud data.

Table 2 queries for crime-related news search

|  |  |
| --- | --- |
| **Group** | **Queries** |
| assault\_query | "assault, "assaults", "assaulter", "assaulters", "assaulted", "assaulting", "assailant", "assailants" |
| crime\_query | "crime, "crimes", "criminal", "criminals" |
| drug\_query | "criminal gang", "criminal gangs", "drug gang", "drug gangs", "gangster", "organised crime", "illegal drug", "illegal drugs", "drug dealer", "drug dealers", "drug dealing", "drug smuggler", "drug smugglers", "drug smuggling", "drug trafficking", "drug trafficker", "drug traffickers", "drug cartel", "drug cartels", "drug peddler", "drug peddlers", "drug peddling", "cocaine", "mdma", "ecstasy", "ketamine", "heroin", "fentanyl", "meth" |
| fraud\_query | "fraud", "frauds", "frauding", "frauded", "fraudster", "fraudsters", "scam", "scammer", "scams", "scammers", "scammed", "scamming" |
| murder\_query | "murder", "murders", "murderer", "murderers", "murdered", "murdering", "murderess", "murderesses", "homicide", "homicides", "stab", "stabs”, “stabbed", "stabbing", "knife crime", "knife-crime", "knife attack", "mugger", "muggers", "mugging", "mugged" |
| sex\_query | "sex offender", "sex offenders", "rape", "raper", "raped", "raping", "rapers", "sex assault", "sexual assault", "sexual predator", "sex predator", "sexual predators", "sex predators", "sex abuse", "sexual abuse", "sex abuser", "sexual abuser", "sex abusers", "sexual abusers", "sexually abused", "sexually abusing", "harassment", "harasser", "harassers", "harassing", "harassed", "domestic violence", "femicide", "feminicide", "domestic abuse", "domestic abuser", "gender violence", "domestic violence", "sexual violence", "sex violence", "sexual victim", "sexual victimisation", "sexual victims", "intimate partner violence" |
| thief\_query | "thief", "thieves", "thieved", "thieving", "theft", "thefts", "burglar", "burglary", "burglars", "burglaries", "robbers", "robbed", "robbing", "robber", "steal", "steals", "stolen", "stealing", "stealer", "stealers", "shoplifting", "shoplifter", "shoplifters", "shoplift", "shoplifted", "pickpocket", "pickpocketing", "pickpockets", "pickpocketed", "pickpocket" |

Once these tweets were downloaded, only those tweets made by the selected users were kept (eliminating retweets), and only those tweets that contained at least one reply or quote were sub-selected. This information is sufficient to ask the API to download all those tweets in response to any previously downloaded tweets and any tweets that quote them. The Twitter interface offers the option to reply to a tweet. In doing so, a new tweet is created, mentioning the user or users who made the tweet that is being replied. Twitter generates an identity for each conversation that occurs through a message, making it possible to track the conversation and download each reply. A very different option is that of a quote. In this scenario, the user can reproduce a third party's tweet on their profile and insert some comment into the original tweet. That is, the user comments and attaches another user's tweet, which is somehow relevant (Garimella et al. 2016).

72,860 tweets from news portals and about 1,336,555 user replies or quotes to these tweets were collected (after eliminating duplicate cases, tweets without text, or exclusively with links, and retweets).

### Data Annotation

Punitive reactions to crime were classified on Twitter by training an automated classifier following a supervised learning strategy (Bi et al. 2019). We used traditional Machine Learning and Transformers techniques (more information in the following section). These techniques required the manual annotation of tweets, which was carried out by one of the project researchers as the lead annotator and three criminology MSc students from the University of Manchester were recruited specifically for this task.

Classifying tweets into punitive or non-punitive is not easy and has certain margins of ambiguity. An annotation guide was created to facilitate agreement among annotators (Annex 1), and the lead annotator had general and personalised interviews with the rest of the annotators.

A punitive tweet was defined as any tweet that expressed support for the increased cost of crime by criminals. This includes support for the death penalty, harsh treatment, longer sentences, poor prison conditions, reduced rehabilitative measures, deportation or expulsion from the country, among others.

Some examples (presented to the annotators in the guide):

* *“They need to bring back d\*ath by firing squad”*
* *“Unconditional bail?????????? Why are yall letting him roam free wtf”*
* *“Lowlife waste of oxygen, throw away the key”*
* *“@user let those damned devils rot in jail. And may the little Angel rest in peace”*
* *“@user @user Keep him in a small, dark, lice-ridden cell”*
* *"@user @user A pretty joke. 60 years would be fair. "*
* *"@user So would the 2w guys not shot the driver if not B lock them up for life"*
* *"Set him on fire"*

In the first instance, the annotators iteratively annotated 2100 tweets. Each annotator worked on the same tweets to determine inter-annotator agreement. The average Kappa score obtained was .74, considered moderate regarding interrater reliability (McHugh 2012).

Once these results were obtained, the annotation task was parallelised, i.e., each annotator received a new set of 2000 tweets to classify, which they performed autonomously. After this a total of 10,000 tweets were annotated. This database was used for the training of the Machine Learning and Transformers models. The final database is highly unbalanced, with only 5% of cases classified as "punitive".

### Automated Classifier Training

#### Traditional Machine Learning Techniques

“Machine Learning" can be defined as any algorithm that allows the computer to "learn" a specific task without necessarily being programmed for it. This learning occurs when the computer optimises its task by improving its performance through “experience”. This experience is gained through iteration of attempts to fit the model to the data, using some fit indicator as a measure of how well it’s doing its job. In this sense, all the tools used in this project can be considered Machine Learning tools (including Transformers) (Bi et al. 2019).

However, it is possible to distinguish between traditional strategies and new natural language processing techniques based on deep learning and transformers. Traditional learning algorithms use manually designed features, while deep learning algorithms automatically learn features from input data without relying on handcrafted features. The unstructured nature of human language creates challenges for automated text classification methods. Traditional techniques struggle with capturing long-term dependencies, which Transformers are good at (Mutanga et al. 2020).

The following traditional Machine Learning models are tested in this project:

* KNN (Guo et al. 2003)
* Logistic Regression (Hosmer Jr et al. 2013)
* Decision Trees (Rokach and Maimon 2005)
* Random Forest (Rigatti 2017)
* Support Vector Machines (Noble 2006)
* XGBoost (Chen and Guestrin 2016)
* AdaBoost (Schapire 2013)

All these techniques require a training set and a model validation set. In all cases, the split was 80%-20% respectively.

These models require pre-processing the textual information, a stage called "feature extraction". This stage involves translating the textual information into a numerical representation that Machine Learning algorithms can interpret. Traditionally, the Bag of Words (BoW) strategy was used for this transformation. The BoW model works by creating a vocabulary of all the unique words in a corpus of text. Then, each document in the corpus is represented as a vector that contains the number of times each word in the vocabulary appears in the document.

In essence, word embedding techniques enable us to represent words from a vocabulary in a lower-dimensional continuous vector space. These methods extract word profiles unsupervised by considering the contexts in which the words appear. Unlike bag-of-words models, word embedding techniques go beyond the limitations of individual words and capture both syntactic and semantic aspects (Enríquez et al. 2016).

Both vector representation strategies for classification were tested in this project. All analyses were performed in Python. For feature extraction using Bag of Words, the Scikit Learn library was used (Pedregosa et al. 2011). First, the documents were tokenised, eliminating user names, links, special characters, and stopwords. All words were lemmatised, following standard procedures. In order to reduce the dimensionality problems in the data, Latent semantic Analysis (LSA) using truncated SVD was implemented over the Bag of Words representation.

On the other hand, the imbalance between the class of interest and the majority class (5%-95%) can be an obstacle to the ability of these models to predict correctly. In order to overcome this obstacle, the models were tested with and without undersampling. Undersampling is a very common technique (Mohammed et al. 2020) that involves reducing the majority class to fit a specific ratio. In this case, the 25%-75% split was tested.

Working with the entire base, 2000 components were maintained, explaining 78.4% of the variance. In the exercises with oversampling, 750 components were maintained, explaining a total of 75.8% of the variance. The preprocessing stage ended with a training base with 7996 rows and 2000 columns for the case without undersampling and 1892 rows and 750 columns for the case with undersampling.

The models were fitted to their hyperparameters in all cases through a cross-validated search on the training set (5-fold) (Berrar 2019). The adjustment indicator used was the f1 score (Goutte and Gaussier 2005). The specific characteristics of the hyperametrization (the search space and the results obtained) can be found in Annex 2 (“Annex\_2\_ML\_models.xlsx”). This process involved training 28 models (7 algorithms \* 2 sampling strategies \* 2 feature engineering strategies).

#### Transformers Models

The Transformers architecture (Vaswani et al. 2017) has quickly become dominant in Natural Language Processing, outperforming alternatives (Wolf et al. 2020). This type of model captures long-range sequence features efficiently and allows their pre-training on large text corpora, a strategy that results in pre-trained models that can later be used in downstream tasks with significant performance gains, for example, in classification tasks (Wolf et al. 2020).

This revolution in natural language processing led to the emergence of a series of publicly available pre-trained models that can be fine-tuned for specific tasks. These models are nucleated in the Hugging Face platform[[1]](#footnote-1) (Jain 2022).

In this project, the Python "simpletransformers"[[2]](#footnote-2) library was used to preprocess and fine-tune transformer models for the automated classification of punitive tweets. The 80-20 split between training and testing set was maintained as with traditional Machine Learning models. Five pre-trained models were tested, each with between 3 and 5 epochs (Brownlee 2018). These models were tested with and without undersampling.

The five models tested were:

* distilbert-base-uncased (Sanh et al. 2019)
* bert-base-uncased (Devlin et al. 2018)
* roberta-base (Liu et al. 2019)
* elecetra-base-discriminator (Clark et al. 2020)
* bertweet-base (Nguyen et al. 2020)

All models are accessible through Hugging Face and can be fine-tuned through the simpletransformers library.

### Model Interpretability

Although the model's predictive capacity shows its usefulness for analysing online punitive attitudes, this type of strategy is considered a "black box" type. It is difficult to understand what elements the model considers to define the classification, obsucring its interpretation.

Local Interpretable Model-Agnostic Explanations (LIME) were used to validate the chosen models. LIME is a method for explaining the predictions of any machine learning model. It works by creating a simplified model that approximates the behaviour of the original model around a specific prediction. This simplified model is then used to explain why the original model made the prediction that it did. It is an agnostic strategy, i.e., it accepts using different types of Machine Learning models. This strategy allows us to account for whether the model follows a human-like logic when classifying (Ribeiro et al. 2016).

## Results

### Traditional Machine Learning Techniques

Despite the plurality of strategies implemented, the results were modest, especially compared to the Transformer models. The highest f1-score obtained on the test set was with the logistic model without undersampling and with word embeddings (.524). Given the poor results compared with the Transformer models, we did not delve into this model type for classification or explanation through LIME. The results of the different models can be found in Annexes 2 and 3.

### Transformers

The transformer models performed, on average better than traditional Machine Learning models. Tests performed with undersampling achieved very high f-score values on balanced test sets; however, they worsened the performance on unbalanced tests, which is what we expect to find in real data in social media, so we chose to continue with unbalanced data (Annexes 3). The most consistent pre-training model was BERTweet (Nguyen et al. 2020), which is to be expected since it is trained on a text corpus similar to the one used in our study (from Twitter).

Once this model was established as the best-performing model, tests were performed on different partitions and initialisations for the BERTweet model between 3 and 5 epochs. Five different initialisations were performed on five partitions, for each of the epochs. The 3-epoch models resulted in a better f1-score value (0.772). The confusion matrix is presented below.

Figure 2 Confusion Matrix



### Inference

The trained classifier was used to predict punitive attitudes on the complete database. The proportion of punitive tweets, total punitive tweets, and total replies by date are shown below (figure 3). These are the measures that will be used for the subsequent time series study of the effects and determinants of online punitive attitudes.

#### Time series

Figure 3 Time series for punitive attitudes and total replies



#### Top cases

In order to validate the tool, the tweets that collected the most punitive responses and the proportion of punitive tweets per query were observed. It is expected that he cases involving the most violent crimes are the ones that accumulate a higher proportion of punitive tweets, as seen in the literature (Tajalli et al. 2013; Aizpurúa and Fernández 2016). This seems to hold true, based on the information below.

Table 3 Top news based on punitive reactions

|  |  |
| --- | --- |
| News Item | Sum Punitives |
| BREAKING: Emma Tustin has been jailed for life with a minimum term of 29 years for starving, poisoning and then murdering her six-year-old stepson Arthur Labinjo-Hughes. Arthur's father Thomas Hughes has been jailed for a minimum of 21 yearshttps://t.co/Dp4sYKZYZx | 716 |
| The 27-year prison sentence for the teenager convicted of abducting, raping and murdering Alesha MacPhail was "excessive and a miscarriage of justice", an appeal court has heard https://t.co/2zdhUsBTYB | 634 |
| Hashem Abedi, convicted of murdering 22 people in the 2017 Manchester Arena attack, refuses to leave cell for sentencing hearing https://t.co/PVSONQtHjU | 600 |
| Sofija Kaczan death: Mugger jailed for killing 100-year-old woman https://t.co/PwlvZLsnO0 | 597 |
| Lee Rigby killer fighting for life in hospital after getting Covid https://t.co/YBUhach0pj https://t.co/GMYNAswoTm | 512 |
| Hunt for thug who wrestled elderly woman, 79, to the ground and broke her arm while stealing her handbag https://t.co/GddUDWOXxY https://t.co/rb75Szr0Yr | 486 |
| Family killer Chris Watts 'back in touch with lover he killed wife and kids for'https://t.co/bACnlmUrH2 https://t.co/hmpSQRSgs8 | 364 |
| Teen who stabbed schoolboy Yousef Makki to death freed from jail after seven monthshttps://t.co/hdMuoN5FG9 https://t.co/AF4hX9fvFB | 362 |
| BREAKING: Thomas Griffiths has been sentenced to life with the minimum of 12 years and six months for murdering his ex-girlfriend Ellie Gould.Get the latest on this story here: https://t.co/DXotFGcaJ7 https://t.co/RdqOTihji2 | 357 |
| Man sentenced to 24 years after sexual assault of 12-day-old baby https://t.co/wBDUidth3w | 324 |

#### Inference by query

Table 4 Punitive proportion by query

|  |  |  |  |
| --- | --- | --- | --- |
| query | total | punitive | proportion |
| assault | 69783 | 3587 | 5.1% |
| crime | 205029 | 5189 | 2.5% |
| drug | 40383 | 1060 | 2.6% |
| fraud | 56945 | 1133 | 2.0% |
| murder | 481811 | 35380 | 7.3% |
| sex | 223297 | 15318 | 6.9% |
| thief | 128211 | 5225 | 4.1% |

### Explanation

The LIME analysis seems to validate the model. A number of theoretically relevant examples were selected to observe the salient elements. Three types of indicators were chosen, commonly associated with punitive attitudes: support for the death penalty, support for harsh treatment, and support for longer sentences or life imprisonment.

#### Death Penalty







#### Harsh treatment







#### Life sentences







## Next steps

The central objective of this project was to create and validate a Machine Learning strategy to capture online punitive reactions. We believe that this has been successfully achieved. The team aims to continue working on the model, since we think that we can improve this initial and promising findings with the goal of develop a pipeline of publications. We plan to do this by refining the hyperparameterization of the models as well as by testing other feature extraction strategies. As a secondary objective, the interdisciplinary work between professionals from the social sciences and computational sciences has also been highly satisfactory, allowing us to identify relevant synergies which will be useful future collaborations in publications and projects.

The following steps involve taking advantage of this measure for the production of academic outputs.

The research group has generated a future research agenda that contemplates the publication of at least three papers:

* A technical paper describing the tool created and its potential. To be published in journals more oriented to computational sciences (e.g. Journal of Computational Social Science).
* A paper highlighting the determinants of punitive attitudes online, involving multilevel analysis to analyse the concrete effects of tweet-, user-, and news-level characteristics. This paper points to criminological journals (e.g. British Journal of Criminology, Criminology, Crime Science).
* A third paper analysing the fluctuations in punitive attitudes and their effects and determinants. Here we expect to use time series models (e.g. ARIMAX or Vector Autoregressive Regressions - VAR). This paper would also target criminological or computational social science-oriented journals (already mentioned).

In addition, these preliminary results are expected to be presented at conferences to be defined by the research team, such as the European Society of Criminology conference, the annual conference of the British Society of Criminology or the Open Data Science Conference.

It is also expected to generate dissemination materials that will allow access to this information to the general public, especially stakeholders at the political and civil society levels.

# References

Adriaenssen, A. and Aertsen, I. 2015. Punitive attitudes: Towards an operationalization to measure individual punitivity in a multidimensional way. *European Journal of Criminology* 12(1), pp. 92–112. doi: 10.1177/1477370814535376.

Aguilar-Jurado, J.A. 2018. Aproximación al análisis de las actitudes punitivas. *Revista Criminalidad* 60(1), pp. 95–110.

Aizpurúa, E. 2015. Delimitando el punitivismo. Las actitudes de los españoles hacia el castigo de los infractores juveniles y adultos. *Revista Española de Investigación Criminológica* 13, pp. 1–30.

Aizpurúa, E. and Fernández, E. 2016. Opinión pública hacia el castigo de los delincuentes: la importancia del delito cometido. *Boletín Criminológico* (22). Available at: https://revistas.uma.es/index.php/boletin-criminologico/article/view/7549 [Accessed: 30 July 2023].

Ajzen, I. and Fishbein, M. 2005. The Influence of Attitudes on Behavior. In: *The handbook of attitudes*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, pp. 173–221.

Albarracin, D. and Shavitt, S. 2018. Attitudes and Attitude Change. *Annual Review of Psychology* 69(1), pp. 299–327. doi: 10.1146/annurev-psych-122216-011911.

Armborst, A. 2017. How fear of crime affects punitive attitudes. *European Journal on Criminal Policy and Research* 23(3), pp. 461–481. doi: 10.1007/s10610-017-9342-5.

Bail, C.A. 2014. The cultural environment: Measuring culture with big data. *Theory and Society* 43(3), pp. 465–482.

Bail, C.A., Brown, T.W. and Mann, M. 2017. Channeling hearts and minds: Advocacy organizations, cognitive-emotional currents, and public conversation. *American Sociological Review* 82(6), pp. 1188–1213.

Barrie, C. and Ho, J.C. 2021. academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint. *Journal of Open Source Software* 6(62), p. 3272.

Berrar, D. 2019. Cross-Validation. In: Ranganathan, S., Gribskov, M., Nakai, K., and Schönbach, C. eds. *Encyclopedia of Bioinformatics and Computational Biology*. Oxford: Academic Press, pp. 542–545. Available at: https://www.sciencedirect.com/science/article/pii/B978012809633820349X [Accessed: 27 July 2023].

Bi, Q., Goodman, K.E., Kaminsky, J. and Lessler, J. 2019. What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology* 188(12), pp. 2222–2239. doi: 10.1093/aje/kwz189.

Boda, Z. and Szabó, G. 2011. The media and attitudes towards crime and the justice system: A qualitative approach. *European journal of criminology* 8(4), pp. 329–342.

Bottoms, A. 1995. The philosophy and politics of punishment and sentencing. *The politics of sentencing reform* 17.

Britannica, T. 2023. *Twitter | Company, History, Description, Elon Musk, & Uses | Britannica*. Available at: https://www.britannica.com/topic/Twitter [Accessed: 16 March 2023].

Brooks, T. 2021. *Punishment: A Critical Introduction*. 2nd ed. London: Routledge. doi: 10.4324/9781315527772.

Brownlee, J. 2018. What is the Difference Between a Batch and an Epoch in a Neural Network. *Machine Learning Mastery* 20.

Buil-Gil, D., Miró-Llinares, F., Moneva, A., Kemp, S. and Díaz-Castaño, N. 2021. Cybercrime and shifts in opportunities during COVID-19: a preliminary analysis in the UK. *European Societies* 23(sup1), pp. S47–S59. doi: 10.1080/14616696.2020.1804973.

Campregher, J. and Jeglic, E.L. 2016. Attitudes Toward Juvenile Sex Offender Legislation: The Influence of Case-Specific Information. *Journal of Child Sexual Abuse* 25(4), pp. 466–482. doi: 10.1080/10538712.2016.1153558.

Chen, T. and Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. New York, NY, USA: Association for Computing Machinery, pp. 785–794. Available at: https://dl.acm.org/doi/10.1145/2939672.2939785 [Accessed: 26 July 2023].

Clark, K., Luong, M.-T., Le, Q.V. and Manning, C.D. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cook, K.J. and Powell, C. 2003. Christianity and punitive mentalities: A qualitative study. *Crime, Law and Social Change* 39(1), pp. 69–89. doi: 10.1023/A:1022487430900.

Cullen, F.T., Jonson, C.L. and Nagin, D.S. 2011. Prisons do not reduce recidivism: The high cost of ignoring science. *The Prison Journal* 91(3\_suppl), pp. 48S-65S.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Edelmann, A., Wolff, T., Montagne, D. and Bail, C.A. 2020. Computational social science and sociology. *Annual Review of Sociology* 46, pp. 61–81.

Enns, P.K. 2014. The Public’s Increasing Punitiveness and Its Influence on Mass Incarceration in the United States. *American Journal of Political Science* 58(4), pp. 857–872. doi: https://doi.org/10.1111/ajps.12098.

Enríquez, F., Troyano, J.A. and López-Solaz, T. 2016. An approach to the use of word embeddings in an opinion classification task. *Expert Systems with Applications* 66, pp. 1–6. doi: 10.1016/j.eswa.2016.09.005.

Flores, R.D. 2017. Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona’s SB 1070 Using Twitter Data. *American Journal of Sociology* 123(2), pp. 333–384. doi: 10.1086/692983.

Garimella, K., Weber, I. and De Choudhury, M. 2016. Quote RTs on Twitter: usage of the new feature for political discourse. In: *Proceedings of the 8th ACM Conference on Web Science*. WebSci ’16. New York, NY, USA: Association for Computing Machinery, pp. 200–204. Available at: https://dl.acm.org/doi/10.1145/2908131.2908170 [Accessed: 7 July 2023].

Garland, D. 2021. What’s wrong with penal populism? Politics, the public, and criminological expertise. *Asian Journal of Criminology* 16(3), pp. 257–277.

Gerbaudo, P. 2014. Populism 2.0: Social media activism, the generic Internet user and interactive direct democracy. In: *Social Media, Politics and the State*. Routledge.

Goutte, C. and Gaussier, E. 2005. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In: Losada, D. E. and Fernández-Luna, J. M. eds. *Advances in Information Retrieval*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 345–359. doi: 10.1007/978-3-540-31865-1\_25.

Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K. 2003. KNN Model-Based Approach in Classification. In: Meersman, R., Tari, Z., and Schmidt, D. C. eds. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 986–996. doi: 10.1007/978-3-540-39964-3\_62.

Haman, M. 2020. The use of Twitter by state leaders and its impact on the public during the COVID-19 pandemic. *Heliyon* 6(11), p. e05540.

Hartman, R. et al. 2022. Interventions to reduce partisan animosity. *Nature Human Behaviour* 6(9), pp. 1194–1205. doi: 10.1038/s41562-022-01442-3.

Himelboim, I., Sweetser, K.D., Tinkham, S.F., Cameron, K., Danelo, M. and West, K. 2016. Valence-based homophily on Twitter: Network Analysis of Emotions and Political Talk in the 2012 Presidential Election. *New Media & Society* 18(7), pp. 1382–1400. doi: 10.1177/1461444814555096.

Hofman, J.M. et al. 2021. Integrating explanation and prediction in computational social science. *Nature* 595(7866), pp. 181–188. doi: 10.1038/s41586-021-03659-0.

Horstman, N.J., Bond, C.E. and Eriksson, L. 2021. Sentencing domestic violence offenders: A vignette study of public perceptions. *Journal of interpersonal violence* 36(21–22), pp. NP11916–NP11939.

Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X. 2013. *Applied logistic regression*. John Wiley & Sons.

Hu, L. and Kearney, M.W. 2021. Gendered Tweets: Computational Text Analysis of Gender Differences in Political Discussion on Twitter. *Journal of Language and Social Psychology* 40(4), pp. 482–503. doi: 10.1177/0261927X20969752.

Jahnke, S. 2018. Emotions and Cognitions Associated with the Stigma of Non-Offending Pedophilia: A Vignette Experiment. *Archives of Sexual Behavior* 47(2), pp. 363–373. doi: 10.1007/s10508-017-1073-7.

Jain, S.M. 2022. Hugging Face. In: Jain, S. M. ed. *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems*. Berkeley, CA: Apress, pp. 51–67. Available at: https://doi.org/10.1007/978-1-4842-8844-3\_4 [Accessed: 27 July 2023].

Jennings, W., Farrall, S., Gray, E. and Hay, C. 2017. Penal Populism and the Public Thermostat: Crime, Public Punitiveness, and Public Policy. *Governance* 30(3), pp. 463–481. doi: 10.1111/gove.12214.

Jin, F., Wang, W., Zhao, L., Dougherty, E., Cao, Y., Lu, C.-T. and Ramakrishnan, N. 2014. Misinformation Propagation in the Age of Twitter. *Computer* 47(12), pp. 90–94. doi: 10.1109/MC.2014.361.

Kirk, I. 2022a. *Britons don’t tend to support the death penalty… until you name the worst crimes | YouGov*. Available at: https://yougov.co.uk/topics/politics/articles-reports/2022/03/30/britons-dont-tend-support-death-penalty-until-you- [Accessed: 12 October 2022].

Kirk, I. 2022b. *Criminal sentencing is too soft, say two-thirds of Britons | YouGov*. Available at: https://yougov.co.uk/topics/politics/articles-reports/2022/03/30/criminal-sentencing-too-soft-say-two-thirds-briton [Accessed: 12 October 2022].

Kostakos, P. 2018. Public perceptions on organised crime, mafia, and terrorism: a big data analysis based on twitter and Google trends. *International Journal of Cyber Criminology* 12(1), pp. 282–299.

KPMG. 2022. *UK economy marred by uncertainty while high inflation takes its toll on global growth prospects - KPMG United Kingdom*. Available at: https://home.kpmg/uk/en/home/media/press-releases/2022/09/uk-economy-marred-by-uncertainty-while-high-inflation.html [Accessed: 12 October 2022].

Leis, A., Ronzano, F., Mayer, M.A., Furlong, L.I. and Sanz, F. 2019. Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. *Journal of Medical Internet Research* 21(6), p. e14199. doi: 10.2196/14199.

Leverentz, A. 2011. Neighborhood context of attitudes toward crime and reentry. *Punishment & Society* 13(1), pp. 64–92.

Liu, P., Nunn, R. and Shambaugh, J. 2018. The economics of bail and pretrial detention. *Economic analysis*.

Liu, Y. et al. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loeffler, C.E. and Nagin, D.S. 2022. The impact of incarceration on recidivism. *Annual Review of Criminology* 5, pp. 133–152.

Maguire, E. and Johnson, D. 2015. The structure of public opinion on crime policy: Evidence from seven Caribbean nations. *Punishment & Society* 17(4), pp. 502–530. doi: 10.1177/1462474515604385.

Mascini, P. and Houtman, D. 2006. Rehabilitation and Repression: Reassessing their Ideological Embeddedness. *The British Journal of Criminology* 46(5), pp. 822–836. doi: 10.1093/bjc/azl014.

McHugh, M.L. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica* 22(3), pp. 276–282.

Mohammed, R., Rawashdeh, J. and Abdullah, M. 2020. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In: *2020 11th International Conference on Information and Communication Systems (ICICS)*. pp. 243–248. doi: 10.1109/ICICS49469.2020.239556.

Mutanga, R.T., Naicker, N. and Olugbara, O.O. 2020. Hate speech detection in twitter using transformer methods. *International Journal of Advanced Computer Science and Applications* 11(9).

Nguyen, D.Q., Vu, T. and Nguyen, A.T. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200*.

Noble, W.S. 2006. What is a support vector machine? *Nature biotechnology* 24(12), pp. 1565–1567.

O’Connor, C.D. 2017. The police on Twitter: image management, community building, and implications for policing in Canada. *Policing and Society* 27(8), pp. 899–912. doi: 10.1080/10439463.2015.1120731.

Office for National Statistics (UK). 2021. *Homicide rate per million population in England and Wales from 2002/03 to 2020/21*. Available at: https://www.statista.com/statistics/318385/homicide-rate-england-and-wales/ [Accessed: 12 October 2022].

Office for National Statistics (UK). 2022. *Number of police recorded violence against the person offences in England and Wales from 2002/03 to 2021/22*. Available at: https://www.statista.com/statistics/288256/violent-crimes-in-england-and-wales/ [Accessed: 12 October 2022].

Ortet-Fabregat, G. and Pérez, J. 1992. An assessment of the attitudes towards crime among professionals in the criminal justice system. *The British Journal of Criminology* 32(2), pp. 193–207.

Ozalp, S., Williams, M.L., Burnap, P., Liu, H. and Mostafa, M. 2020. Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech. *Social Media + Society* 6(2), p. 2056305120916850. doi: 10.1177/2056305120916850.

Pedregosa, F. et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12(85), pp. 2825–2830.

Piquero, A.R., Cullen, F.T., Unnever, J.D., Piquero, N.L. and Gordon, J.A. 2010. Never too late: Public optimism about juvenile rehabilitation. *Punishment & Society* 12(2), pp. 187–207. doi: 10.1177/1462474509357379.

Powell, A., Overington, C. and Hamilton, G. 2018. Following# JillMeagher: Collective meaning-making in response to crime events via social media. *Crime, Media, Culture* 14(3), pp. 409–428.

Prabhakar Kaila, D.R. and Prasad, D.A.V.K. 2020. Informational Flow on Twitter – Corona Virus Outbreak – Topic Modelling Approach. Available at: https://papers.ssrn.com/abstract=3565169 [Accessed: 15 August 2022].

Prieto Curiel, R., Cresci, S., Muntean, C.I. and Bishop, S.R. 2020. Crime and its fear in social media. *Palgrave Communications* 6(1), pp. 1–12. doi: 10.1057/s41599-020-0430-7.

Radanliev, P. and De Roure, D. 2023. New and emerging forms of data and technologies: literature and bibliometric review. *Multimedia Tools and Applications* 82(2), pp. 2887–2911. doi: 10.1007/s11042-022-13451-5.

Ramirez, M.D. 2013. Punitive Sentiment. *Criminology* 51(2), pp. 329–364. doi: https://doi.org/10.1111/1745-9125.12007.

Ribeiro, M.T., Singh, S. and Guestrin, C. 2016. Model-Agnostic Interpretability of Machine Learning. Available at: http://arxiv.org/abs/1606.05386 [Accessed: 27 July 2023].

Riemer, H., Shavitt, S., Koo, M. and Markus, H.R. 2014. Preferences don’t have to be personal: Expanding attitude theorizing with a cross-cultural perspective. *Psychological Review* 121(4), pp. 619–648. doi: 10.1037/a0037666.

Rigatti, S.J. 2017. Random Forest. *Journal of Insurance Medicine* 47(1), pp. 31–39. doi: 10.17849/insm-47-01-31-39.1.

Rokach, L. and Maimon, O. 2005. Decision Trees. In: Maimon, O. and Rokach, L. eds. *Data Mining and Knowledge Discovery Handbook*. Boston, MA: Springer US, pp. 165–192. Available at: https://doi.org/10.1007/0-387-25465-X\_9 [Accessed: 26 July 2023].

Rufai, S.R. and Bunce, C. 2020. World leaders’ usage of Twitter in response to the COVID-19 pandemic: a content analysis. *Journal of public health* 42(3), pp. 510–516.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schapire, R.E. 2013. Explaining AdaBoost. In: Schölkopf, B., Luo, Z., and Vovk, V. eds. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*. Berlin, Heidelberg: Springer, pp. 37–52. Available at: https://doi.org/10.1007/978-3-642-41136-6\_5 [Accessed: 26 July 2023].

Singh, L., Bode, L., Budak, C., Kawintiranon, K., Padden, C. and Vraga, E. 2020. Understanding high- and low-quality URL Sharing on COVID-19 Twitter streams. *Journal of Computational Social Science* 3(2), pp. 343–366. doi: 10.1007/s42001-020-00093-6.

Socia, K.M., Rydberg, J. and Dum, C.P. 2019. Punitive Attitudes Toward Individuals Convicted of Sex Offenses: A Vignette Study. *Justice Quarterly* 0(0), pp. 1–28. doi: 10.1080/07418825.2019.1683218.

Stalans, L.J. 2013. Measuring attitudes to sentencing. In: *Changing attitudes to punishment*. Willan, pp. 33–50.

Stephens-Davidowitz, S. 2018. *Everybody lies: What the internet can tell us about who we really are*. Bloomsbury Publishing.

Tajalli, H., De Soto, W. and Dozier, A. 2013. Determinants of Punitive Attitudes of College Students Toward Criminal Offenders. *Journal of Criminal Justice Education* 24(3), pp. 339–356. doi: 10.1080/10511253.2012.740055.

Vaswani, A. et al. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Williams, M.L. and Burnap, P. 2016. Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. *The British Journal of Criminology* 56(2), pp. 211–238. doi: 10.1093/bjc/azv059.

Williams, M.L., Burnap, P., Javed, A., Liu, H. and Ozalp, S. 2020. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology* 60(1), pp. 93–117. doi: 10.1093/bjc/azz049.

Wolf, T. et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. Available at: https://aclanthology.org/2020.emnlp-demos.6 [Accessed: 27 July 2023].

Wood, J.L. and Viki, G.T. 2001. Public attitudes to crime and punishment: A review of the research. *Report prepared for the Esmee Fairbairn Charitable Trust. UK.*

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y. and Zhu, T. 2020. Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach. *Journal of medical Internet research* 22(11), p. e20550.

Yang, K.-C., Torres-Lugo, C. and Menczer, F. 2020. Prevalence of low-credibility information on twitter during the covid-19 outbreak. *arXiv preprint arXiv:2004.14484*.

1. https://huggingface.co/ [↑](#footnote-ref-1)
2. https://github.com/ThilinaRajapakse/SimpleTransformers [↑](#footnote-ref-2)