Classifying Hate: Legal and Ethical Evaluations of ML-Assisted Hate Crime Classification and Estimation in Sweden

Holli Sargeant
University of Cambridge
Cambridge, United Kingdom
hs775@cam.ac.uk

Hannes Waldetoft
Uppsala University
Uppsala, Sweden
hannes.waldetoft@statistik.uu.se

Måns Magnusson
Uppsala University
Uppsala, Sweden
mans.magnusson@statistik.uu.se

Abstract

Hate crimes, driven by biases against specific demographic groups, harm not only individuals but undermine the security, trust, and cohesion of entire communities. Accurately identifying such crimes remains a significant challenge due to under-reporting, limited training, and the complexity of determining bias motivations. In this paper, we analyze the results of a text classification model developed to improve the precision of hate crime statistics and identification in Sweden. Empirical results indicate the model outperforms traditional manual police classification of hate crimes, achieving higher precision across various crime types and regions. We further disaggregate performance to pinpoint persistent challenges and highlight categories where both human and machine decision-makers struggle. While the model focuses on statistical estimation rather than direct case-level decision-making, we discuss the broader implications of algorithmic transparency, accountability, and explainability. Ultimately, this research illustrates how transformer-based neural networks can responsibly bolster the detection and understanding of hate crimes, informing policies to better protect vulnerable communities.

Content warning: This article includes direct quotations and descriptions from hate crime reports, containing offensive language, hateful symbols, and references to discriminatory actions.

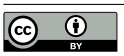
CCS Concepts

• **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → **Machine learning**; **Natural language processing**; • **Mathematics of computing** → **Probability and statistics**.

Keywords

Hate Crime Classification, Automated Crime Classification, Algorithmic Bias, Criminal Justice, Responsible AI, Procedural Justice

ACM Reference Format:
Holli Sargeant, Hannes Waldetoft, and Måns Magnusson. 2025. Classifying Hate: Legal and Ethical Evaluations of ML-Assisted Hate Crime Classification and Estimation in Sweden. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3715275.3732016



FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732016

1 Introduction

Concerns about unfairness, irresponsibility, and opacity are important in most applications of AI but are profoundly significant when applied to the criminal justice system. The stakes are particularly high where such tools can influence decisions that materially affect people's lives, liberty, and rights. A wealth of research has scrutinized AI applications within the criminal justice system, highlighting the potential for these technologies to perpetuate or even exacerbate existing inequalities [1, 7, 17, 21, 97, 98]. Large-scale failures have demonstrated systematic inaccuracies and biases in various algorithmic and predictive tools from predictive policing to individual criminal risk assessments [22, 30, 35, 50, 53, 58, 70, 98]. Despite these concerns, the rapid adoption of such tools by law enforcement has continued, driven by a myriad of factors, including the surge in AI capabilities, declining funding for law enforcement and the justice system, and recent public scrutiny on the role of policing [1]. It reflects an ongoing search for solutions that can overcome resource limitations while delivering greater consistency, impartiality, and efficiency in decision-making processes.

Hate crimes are criminal offenses committed with a bias motivation. These acts, driven by fear, prejudice, or hatred towards specific demographic groups, not only harm individual victims but also impact entire communities with which the victim identifies [48, 64–66]. Accurate classification of whether a crime was bias-motivated is essential for effective law enforcement, allocation of resources, and development of policies to prevent these offenses. However, hate crimes are often under-reported and under-classified for various reasons, including victims' reluctance to report, lack of awareness among law enforcement, and complexities of determining bias-motivation [48, 64]. These challenges warrant investigation into innovative methods that can enhance the accuracy and consistency of hate crime identification.

The Swedish Brottsförebyggande rådet (**Brå**) [National Council for Crime Prevention] plays a central role in compiling, analyzing, and publishing public statistics on crime, including hate crimes. As part of an ongoing effort to improve the identification and classification of hate crimes, our research team has formed an academic collaboration with Brå to explore how transformer-based neural networks can be used in statistical estimation. Through this partnership, we seek to complement and strengthen Brå's existing procedures by leveraging machine learning tools that can assist analysts and policymakers in achieving more reliable, timely, and well-informed hate crime statistics.

In this paper, we aim to evaluate a *responsible and efficient* use of AI in the criminal justice system by leveraging text classifiers to improve hate crime classification in Sweden. Building on our concurrent work, which focused on statistical methods for estimating

annual hate crime statistics [93], we extend the analysis to explore the statistical, legal, and ethical implications of deploying such text classification methods for hate crime classification and estimation. Initially, this work was aimed at providing efficient estimates of yearly hate crimes statistics from the highly imbalanced textual data of police reports, while improving accuracy and minimizing manual annotations. We now examine the broader implications that would arise if this technique was deployed for both estimation or classification of police reports, either by Brå or police. This preemptive approach aims to anticipate and mitigate potential challenges, ensuring that our methods are not only statistically sound but also ethically responsible and legally compliant.

Considerable research has been dedicated to detecting hate speech, particularly focusing on overtly abusive language found on social media platforms and other digital channels [37, 56, 67, 72, 80]. However, our work differs from these studies in both objective and context. We specifically classify official police descriptions of events rather than identifying direct hateful language in user-generated content. Additionally, many of these events involve hate-motivated actions rather than direct speech.

To our knowledge, this is the first study to explore the legal and ethical implementation of machine learning models for hate crime classification. This paper represents a significant advancement in developing innovative methods that demonstrate improved accuracy, efficiency and consistency in crime classifications, while mitigating legal and ethical concerns. Our paper aims to answer three research questions (**RQs**):

- **RQ1:** How does police annotators compare to that of text classification models?
- **RQ2:** Are there systematic biases with respect to crime code, police region, or hate crime motivation?
- **RQ3:** What are the potential legal and ethical implications of deploying this type of model in the criminal justice system?

Our primary contributions include:

- *Text Classification Comparisons:* We evaluate a transformer model adapted to Swedish police reports, comparing it to existing manual police annotations.
- *Analysis of Misclassification Factors:* We identify specific sources of misclassification in police and model decisions, highlighting how nuanced contexts, data sparsity, and nuances in vocabulary affect model performance.
- *Legal and Ethical Implications:* We explore the legal and ethical implications of deploying such models in the justice system, emphasizing the need for robust safeguards even for highly accurate models.

The paper is structured as follows. Section 2 introduces the broader context and legal framework surrounding hate crimes in Sweden. Section 3 details our data collection, concurrent work on model development, and methods used for assessment. Sections 4.1 and 4.2 present our findings, focusing on performance comparisons and systematic biases. Section 4.3 discusses the legal and ethical considerations of deploying such a model in practice. Section 5 concludes.

2 Hate Crimes in Sweden

Sweden's population with a non-European background has increased markedly in recent decades, the populist far-right has expanded its electoral base, and a concurrent increase in racial and xenophobia-motivated crimes has made policing hate crimes a salient issue [4, 6]. Despite Sweden's prioritization of combating hate crimes since the mid-1990s, ongoing challenges persist [38–40]. Sweden has participated in the United Nations Human Rights Council's Universal Periodic Review (**UPR**) three times. In the 2010 and 2015 UPR, Sweden was criticized for the situation concerning hate speech, bias-motivated crimes, and xenophobia and racism, particularly hatred against Islam and towards Muslims [23, 24]. In 2020, the UPR Working Group acknowledged considerable progress but made sixty-one recommendations to combat racism and hate crimes [25]. There is considerable need for enhanced knowledge and more effective mechanisms to identify, investigate, and prosecute these offenses.

2.1 Legal framework for hate crimes

Sweden has four fundamental laws which make up the Constitution: Instrument of Government, Act of Succession, Freedom of Press Act and Fundamental Law on Freedom of Expression. Together with Sweden's ratification and incorporation of international rights conventions, these laws protect free speech, expression, assembly, and association [26, 83, 84]. These freedoms may be limited, especially to prevent harm or hateful speech. Such limitations are exemplified in Sweden's legal provisions addressing hate crimes, which include three specific offenses and a sentencing enhancement: (1) agitation against a population group [82, Ch 16, § 8]; (2) unlawful discrimination [82, Ch 16, § 9]; (3) defamatory crime of insulting behavior [82, Ch 5, § 3]; and (4) a penalty enhancement provision that designates hate motives as aggravating factors during sentencing for most criminal offenses [82, Ch 29, § 2(7)]. This provision was introduced in 1994 to combat the rise of neo-Nazism and to strengthen compliance with the International Convention on the Elimination of All Forms of Racial Discrimination [85]. Since then, the provision has expanded to protect sexual orientation [86], and transgender identity or expression [88]. The penalty enhancement provision states:

> As aggravating circumstances when assessing penalty value, in addition to what applies for each specific type of offence, particular consideration is given to: . . . 7. whether a motive for the offence was to insult a person or a population group on grounds of race, colour, national or ethnic origin, religious belief, sexual orientation or transgender identity or expression, or another similar circumstance; [82, Ch 29, § 2(7)].

Hate motives need not be the sole or primary motive behind the offense; it may be one of several motivating factors [2]. Policy documents and guidance has been issued by Polismyndigheten [Police Authority], Åklagarmyndigheten [Prosecution Authority], and Brå to aid actors in identifying, investigating, prosecuting, and sentencing hate crimes.

2.2 Hate crime assessment

Police rarely investigate motive in sufficient detail to uncover the true reasons for a crime, meaning the bias motive of a hate crime perpetrator may be difficult to identify. Without mechanisms to identify, record, and investigate the bias motive, hate crimes risk remaining undetected [64]. In the early 2000s, Stockholm police started evaluating the use of a pop-up window when recording criminal incidents to flag potential hate crimes [43, 44]. In these early trials, the results showed "that there were no uniform practices in how the police officers categorized crimes as hate crimes" [48]. In 2008, Polismyndigheten commissioned all Swedish police authorities to introduce a marking for each police report as a suspected hate crime or not, and also began developing routines and guidelines to improve hate crime categorization and registration [48, 71]. Since 2019, a pop-up reminds police offers to identify whether a potential hate crime motive was present before submitting police reports, including a short explanation of what constitutes a hate crime [40].

The accuracy of police annotations of bias motive is low. It is widely accepted that police-recorded crime data is often flawed [14]. The legal status of hate crime as an umbrella term for several types of incidents and the corresponding lack of a dedicated crime code within the Swedish offense reporting system makes it difficult for police officers to make consistent interpretations of the hate crime label. The ambiguity and complexity of hate crime classification amplifies these concerns and has been identified as a source of confusion in law enforcement agencies and error in national hate crime statistics [63]. Hagerlid and Granström [40] identified several studies that revealed incorrect interpretations of hate crime policies and guidelines. For example, Atak [4] found that Stockholm police officers believed the bias motive needed to be very clear and the single motivating factor to meet the hate crime classification, which is more restrictive than legal guidance [2]. Front-end officers and investigators have been shown to lack basic knowledge and skills regarding hate crime classification [38]. These inconsistencies can lead to significant discrepancies in hate crime statistics and, more crucially, in the proper handling and prosecution of these offenses.

2.3 Hate crime statistics

Brå is commissioned by the Swedish Government to produce hate crime statistics. Given the lack of accuracy of the police reports, Brå undertakes its own annotations of those flagged incidents and assigns hate crime experts to reannotate the police reports. In 2009, Brå published the results of its first comparison between its reannotated reports and the classifications made by Polismyndigheten: only classifications in 5% of cases were the same between the two authorities [10]. Since then, large educational efforts have attempted to improve the understanding of the hate crime label [40, 48]. In 2020 and 2022, police tagged 6300 and 4800 reports as hate crimes, respectively; Brå concluded that 54% and 56% of those police reports were actually hate crimes [11, 54]. While police accuracy has significancy improved over time, it remains insufficient for Brå to solely rely on police report classifications for hate crimes statistics.

The accuracy of "official statistics" is mandated by Swedish law [89], although hate crime statistics do not have such accreditation. Obtaining official statistics accreditation in Sweden is crucial because it guarantees the objectivity, quality, and accessibility of data that inform public understanding, policy decisions, and government decision-making [89]. One reason is the variability in how hate crimes are identified and recorded [48], coupled with evolving methodologies. There is an ongoing effort to refine these processes with the aim of fulfilling the requirements for official statistics accreditation.[1] Challenges in crime statistics exist internationally; for example, UK police recorded crime data lost its official accreditation in 2014 due to concerns about the integrity and quality of reports [14, 42, 92]. Even without official status, hate crime statistics serve important purposes, contributing valuable knowledge to research, fulfilling international obligations to report hate crimes, and providing a foundation for prevention, identification, and prosecution of these offenses [54]. Manual annotations for hate crimes is time-consuming. Without being able to rely on the police's initial hate crime annotation, Brå experts must review tens of thousands of police reports and given the complexity of hate crime classification, discuss difficult cases among multiple expert annotators. Therefore, current methods depends fully on the police annotations.

Hate crime laws exist in several countries, primarily across Northern America and Europe, but these laws "differ remarkably from country to country with respect to the specification of protected groups, treatment of hate speech, legal standards for establishing bias motivation, and utilization of hate crime statutes in criminal prosecutions" [79]. Despite differences in legal frameworks, many of the challenges of identifying and enforcing hate crime laws identified here "are not unique to Sweden, with international research instead showing that these are pervasive challenges that have come to characterize the implementation of hate crime law in many countries" [see 40, citing [15, 41, 49, 55, 94, 95]]. The need to overcome these challenges has international importance.

2.4 ML-based hate crime classification

In concurrent work, we developed a machine learning model for classifying hate crimes in Swedish police reports [93]. Our approach does not use general-purpose large language models (**LLMs**) due to the sensitive data-constraints and the lack of compute infrastructures for Brå to run LLMs locally. We briefly summarize the model development and results. First, a Swedish roBERTa model [57] was adapted to the specific linguistic domain of police reports through additional unsupervised pre-training. Second, the model was fine-tuned on a supervised classification task that distinguishes whether a police report has a hate-crime motive or not.

Once trained, the model was evaluated on the held-out test set of 2022 police reports, comparing its classification outputs to ground-truth labels from Brå experts. Only a subset of the 2022 data had verified hate-crime annotations, so the F1-score was estimated using random sampling. In the Brå-labeled subset of 2022 reports, the model reached an F1-score exceeding 90%. Extrapolating to the full population of 2022 reports, the results show that a transformer model is better at classifying hate crime than manual police annotations [see results in 93].

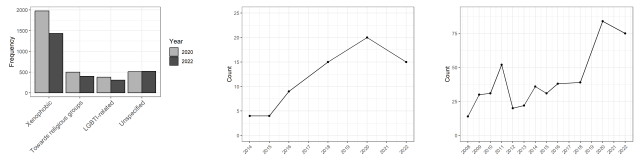
[1] Discussions with Brå Hate Crimes Division (December 2024).

3 Data and Methods

3.1 Data

Available data consists of all Swedish police reports filed between 2007-2022 and was supplied by Brå. In addition to the text of the police report describing the crime, data contains the crime code, municipality code, and police region codes. For the police reports determined by Brå to contain at least one hate crime motive, there is also information on the type of hate crime, the relationship status between the victim and perpetrator, and the type of environment in which the potential crime took place (e.g., domestic, public space, online). We regard these Brå annotations as "gold standard". Brå's hate crime annotators are subject-matter experts with relevant higher-education (i.e., degrees in law, criminology, sociology), have extensive annotation guidelines, formal processes for reviewing police reports and collectively resolving more uncertain or complex police reports.

In total, there are 21.6 million police reports, of which 52,000 are determined to be hate crimes and 61,000 were classified as not having a hate crime motive by Brå experts. The type of hate crime is categorized into four high-level categories (Figure 1a), and then with further subdivisions (Figure 1b and 1c). The subcategories have changed over time, because of changes in law and changes in methodology. As mentioned above, sexual orientation was added to the hate crime definition in 2002 and transphobia was added in 2008. Brå have changed methodology in recording subcategories for several reasons; for example, prior to 2014 anti-Sámi hate crimes were recorded as xenophobia but it is now recorded separately to increase focus on the issue. In 2020, a major revision of the methodology was made, in which the sample for expert annotation was based on the initial classification made by the police instead of a keyword search, so the statistics from 2020 and after are not directly comparable with the preceding years.



(a) Totals (b) Anti-Sámi (c) Transphobic

Figure 1: (a) Bar chart of confirmed hate crimes in 2020 versus 2022, grouped into four main categories: xenophobic, religious, LGBT-related, and other. (b) Timeline of confirmed anti-Sámi hate crimes. (c) Timeline of confirmed transphobic hate crimes. Note that the methodology changed in 2020, so comparisons across the two time periods should be made with caution.

3.2 Bias and error analysis

We have established that the text classification model appears to outperforms police classifications [93]. This paper aims to evaluate the performance disparities between the model, police classifications, and Brå ground-truth annotations. By examining specific instances of misclassification and systematic biases, we seek to uncover the contextual or structural factors that cause either the police or model to deviate from Brå's expert annotations.

In Section 4.1, we assess these discrepancies by reviewing a sample of police reports to identify patterns contributing to misclassifications or inconsistencies. The subsets of police reports, outlined in Table 1, fall into four categories: (1) reports that the model classifies as hate crimes, but the police do not, (2) reports classified by the model as not hate crimes that are mostly accurate according to Brå annotations, (3) reports the model inaccurately classifies as hate crimes, (4) reports the model inaccurately classifies as not hate crimes. From each of these four categories, we randomly sampled police reports for in-depth analysis to better understand the nature of these discrepancies and the limitations of the model in specific contexts. We excluded two scenarios where no discrepancies occur: where Brå, police, and model all classify a report as a hate crime, and where neither the police nor the model classifies a report as a hate crime.

In Section 4.2, we examine the distribution of classification errors systematically across three dimensions: crime codes, police regions, and hate crime motivation categories. By identifying whether certain categories systematically experience higher rates of misclassification, similar to approaches in algorithmic fairness literature [47, 59], we aim to detect potential disparities that may reinforce or exacerbate existing biases. Additionally, we fit a logistic regression in which the outcome y is correct or incorrect classification relative to Brå's ground-truth labels (Appendix B).

4 Results and Discussion

4.1 RQ1: How does police annotators compare to that of text classification models?

To evaluate the text classification model, we analyzed a random sample of police reports to compare its performance to the police and Brå annotations in the different categories outlined in Table 1.

4.1.1 Category 1: Which police reports are classified as hate crimes by the model but not classified as hate crimes by the police? We manually reviewed a random sample of 100 police reports from the 6,568 police reports in this category where the police did not tag reports as hate crimes but the model did classify as hate crimes. Since Brå only reviews police reports that are flagged as potential hate crimes, there are no Brå annotations for these reports. In our evaluation, we noted our own classification of whether they would likely be true or false hate crimes. In our view, 35 police reports would be hate crimes, 52 would not be hate crimes, and 13 were unclear. We aim to understand the instances where the model identifies potential hate crimes that may have been overlooked by the police and why they may have been missed.

Several reports contain indicators of hate that the model likely identified. For example, at least 50 of the reports contained homophobic, racial, and anti-Semitic slurs. The model relies on slurs to classify incidents as hate crimes. An illustrative flagged report contained the word "dumskalle", which translates to "dumbhead". Although "dumskalle" itself is not a slur, the suffix "-skalle" can be combined with other adjectives to form slurs in Swedish. For instance, "svartskalle" ("blackhead") is a derogatory term used to

Table 1: Categories for Review: The table illustrates how the model's classifications compare to the police and Brå classifications. "HC" refers to a hate crime label, and "Not HC" refers to a not hate crime label.

		Police & Brå Classification		
		Police: Not HC	**Police: HC & Brå: HC**	**Police: HC & Brå: Not HC**
Model Classification	**HC**	**Category 1** No Brå annotations for non-flagged reports.	*True Positive* Accurate classification, not analysed further.	**Category 3** *False Negative* Inaccurate classification as HC.
	Not HC	*N/A* No Brå annotations for non-flagged reports.	**Category 4** *False Positive* Inaccurate classification as not HC.	**Category 2** *True Negative** Mostly accurate classification. *Brå annotate 19/20 not HC.

target dark-skinned individuals, especially immigrants. Another report was flagged by the model that included the term "svart" in isolation. In our view, neither of these reports would be classified as hate crimes. Examination of the transformer model's vocabulary revealed that ##skalle is present and consequently, the model may partially recognize the "-skalle" in association with hateful or bias-related language. To test the effect of including the suffix "-skalle", the police report containing "dumskalle" was reclassified with the suffix removed, while keeping all other tokens unchanged. The prediction changed from *hate crime* (estimated probability = 0.64) to *not hate crime* (estimated probability = 0.95). This suggests that including "-skalle", even in a non-slur context, increases the likelihood of classification as a hate crime.

In this case, the model erred by inferring potential hate speech where none existed. This example underscores how sub-word tokenization can contribute to false positives. Additionally, graffiti-related incidents are overrepresented among the reports flagged by the model. In particular, certain words or symbols associated with graffiti and vandalism included terms like "hate", "JEW", and the swastika symbol. The model also identified threats against minority groups and dissemination of extremist materials. It is unclear why these incidents might have been missed by the police. Although we do not have Brå's annotations, it is highly likely that several of these reports should have been flagged as hate crimes based on the presence of hate-related language and symbols.

4.1.2 Category 2: Which police reports are classified as not hate crimes by the model and are mostly accurate according to Brå's expert annotations? We manually reviewed a sample of 50 police reports from the 1,938 instances that the police tagged as hate crimes but the model classified as not hate crimes. Brå annotations confirm that 46 out of 50 police reports were not hate crimes in agreement with the model meaning these are true negatives (and police false positives). These reports related to threats, assaults, graffiti, vandalism, and property damage but lacked definitive bias-related indicators. In many cases, the police may have been influenced by surface-level cues such as the presence of political references to "Antifa", "long live PKK" (Kurdistan Worker's Party) and the Swedish Democrats, that were included in these reports. Many of these police reports were graffiti, vandalism or other property damage where police may have interpreted certain elements, such as symbols or expressions, as hate motivation. In instances where overt hate-related terms were absent, the model consistently refrained from labeling the incident as bias-motivated. Some reports included minimal or vague textual descriptions, providing insufficient evidence for a hate classification. The review of these true negatives highlights that the model not only provides a more precise baseline for detecting hate crimes in police reports but also effectively filter non-hate cases within the pool of reports initially tagged as hate crimes. Such improved initial filtering could streamline annotation procedure, allowing human annotators to focus their efforts on cases more likely to involve genuine bias motivations.

4.1.3 Category 3: Which police reports are classified as hate crimes by both the police and the model but are not classified as hate crimes by Brå's expert annotators? We manually reviewed a sample of 50 police reports that were identified by both the police and the model as hate crimes, which Brå's expert annotators concluded were not hate crimes. In 2022, the police flagged 2,093 false positive reports. Generally, these police reports reveal that the police and model consistently predict hate crimes in cases involving explicit and derogatory language, provocative acts, and specific keywords. Such tendency to over-classify certain incidents as hate crimes is potentially due to the reliance on inflammatory language in threats, defamation, and harassment. Certain words and expressions such as "hitler", "terror", "nazi", "allah akbar", "immigrant", "nigger", and "fucking lesbian bitch" appear to trigger these classifications, even when context or legal definitions may not support a hate crime designation. In one example, harassment of disabled persons, including using the terms "faggot" and "whore", was flagged by police as a hate crime. While the targeted derogatory language aligns with hate crime patterns, disability is not included as a protected characteristic under legal definitions of hate motives, leading to Brå's exclusion of this case. Interestingly, one police reports included the text "might be hate crime"; it is likely this resulted in the incorrect classification by the model.

False Positive Example: Burning the Quran. Out of 50 of these police reports, 12 involved the burning of the Quran. Both the police and model consistently flagged these as hate crimes, but Brå's annotators concluded otherwise. It suggests a systematic discrepancy in interpreting actions targeting religious texts. There has been a significant number of protests in Sweden involving the burning or destruction of the Quran [9, 90]. It is still likely that this number is overrepresented in the data, given that each incident could generate multiple police reports. At face value, it might seem surprising that not only has Brå confirmed these incidents are not hate crimes,

but police have repeatedly granted permission for these protests. It is not that these incidents have not raised legal and security concerns [9, 69]; rather, it highlights the strength of freedom of expression protections under Swedish law [76]. Sweden's Constitution contains strong protections for the freedom of expression. In Sweden, "freedom of expression thus has a peculiar role as a superior human right. In legal cases, there is often a presumption in favour of protecting freedom of expression over other interests or values–such as privacy or honour" [76]. In the context of Quran burning, the offense of agitation against a population group has been considered but, in many cases, these protests are understood as expressing opinions on the Islamic religion not as an attack on the Muslim population as a group [82, Ch 16, § 8]. However, if a protest includes additional hatred or incitement towards a racial or ethnic group it is likely to rise to the hate crime definition.

4.1.4 Category 4: Which police reports are misclassified by the model as not hate crimes but are classified as hate crimes by Brå's expert annotators? We evaluated 50 police reports where the model failed to classify certain police reports as hate crimes, despite Brå's expert annotators finding evidence of bias. In 2022, the police flagged 4,800 reports as hate crimes. Brå identified 2,784 of those reports as police false negatives, and we identified 195 of those reports as model false negatives.

Nonetheless, these 195 model misclassifications highlight areas where the model's detection capabilities can be enhanced, particularly in capturing subtle signals of hate motivation that it currently overlooks. Three of these police reports involved hate crimes targeting the Sámi.

False Negative Example: Hate crimes against Sámi. The Sámi are an indigenous people of northern Scandinavia and Russia, and one of Sweden's five national minorities. The Sámi culture and lifestyle is diverse, encompassing groupings such as the forest Sámi, coastal Sámi, mountain Sámi, Skolt Sámi and Kola Peninsula Sámi [36]. Even the Sámi language has existed in many varieties [61]. The dominant public perception that Sámi culture is uniform and practiced primarily by nomadic reindeer herders is a misconception [36]. From the 17th century onward, state and church policies imposed restrictions on the Sámi's freedom of religion and movement [61]. By the 20th century, existing Sámi resistance to the state was strengthened in response to racial biology and forced assimilation [61]. In 1977, the Sámi gained the status of an indigenous people, and in 2011, they were recognized in the Swedish Constitution [83, Ch 1, § 2]. While there is no precise figure for the Sámi population, the Sámi Parliament's electoral roll had 9,200 registered voters in 2021, and the population is Sweden is commonly estimated as between 20,000 to 40,000 people [91].

Hate crimes against the Sámi occur in many different environments with varying severity. Police reports frequently describe racist insults and threats, as well as crimes related to Sámi reindeer herding, including property vandalism and the killing of reindeer [12]. These incidents often represent retaliatory acts against Sámi rights, and it is common for victims to know the perpetrators, as such offenses frequently occur in smaller communities [12].

We identify three primary reason for these model false negatives on anti-Sámi hate crimes. First, there is only a small number of anti-Sámi hate crimes in our training sets (67 recorded between 2014–2022). Before 2014 hate crimes against the Sámi were not coded separately [12], any prior incidents may have been recorded in the general category of xenophobia. This under-representation likely hinders the model's ability to learn and recognize patterns specific to hate crimes against the Sámi. Second, some of the most common phrases used against the Sámi are not hateful slurs in isolation; the context in which they are used is often needed. The term "Lapp" translates to "patch" or "note" in Swedish, but was originally used to describe the Sámi people's traditional patchwork clothing and the region "Lappland". When used to refer to or about the Sámi people, it can be viewed as derogatory. It is more frequently used in conjunction with other offensive terms; Brå reports that " 'Lappjävel' ['Lapp bastard'] is by far the most common phrase used according to interview participants and police reports", as well as "Lapp-" or "same-" used alongside other slurs [12]. Another example is the term "renknullare", which translates to reindeer [ren] -fucker [-knullare], is sometimes used as an expression aimed at Sámi men [12]. These terms are difficult to capture the extent of context that renders these derogatory or hateful expressions against the Sámi. Third, beyond hateful language, a common and unique form of hate crime against the Sámi involves the deliberate targeting of reindeer in Sámi areas. Perpetrators may attack reindeer, believing that the Sámi's reindeer herding rights limit their own land use [12, 68]. Such actions are significant as they directly impact the Sámi's traditional livelihood and cultural practices. The nature, context, and language in these hate crimes are unique from many other hate crime motivations, and are likely challenging for the model.

4.1.5 Summary of classification issues.

Police classification issues. Police officers may exhibit variability in their classifications due to individual differences in experience, training, personal judgment, and cognitive bias. However, the police may exhibit higher validity due to their ability to interpret complex contexts, nuances, and implicit meanings within their interactions. Several factors contribute to the high rates of classification errors in police hate crime annotations.

First, errors may arise from human error, including simple mistakes, bad report writing habits, and software errors [62]. There are approximately 1.5 million police reports filed every year in Sweden, it can be expected that some reports are mistakenly flagged. Also police may lack adequate knowledge to correctly classify this ambiguous category of crimes [63]. Research indicates some Swedish police have often adopted an overly narrow approach to hate crime classification [4, 40].

Second, conscious or unconscious biases among police officers may influence both the decision to report an incident as bias-motivated and the language used to describe the details in their reports. For example, officers may unintentionally downplay the significance of certain incidents due to personal prejudices, societal stereotypes, or lack of awareness of the impact on minority communities leading to under-classification of hate crimes [4, 29]. Such biases can affect the accuracy and consistency of hate crime reporting, potentially obscuring the true prevalence of these offenses.

Third, political pressures may impact police classifications of hate crimes. Officers might face external influences regarding when to enforce hate crime statutes, affecting their decision-making processes [4, 5, 52]. For instance, Jacobs and Potter [46] argue that

police officers' "unique responsibility for deciding whether particular crimes ought to be labeled 'bias related'... complicates and contributes to the politicization of police operations". Such politicization may result in some officers deliberately downgrading particular crimes to reduce the hate crime rate [62].

While human annotators bring valuable contextual understanding and judgment to the task of hate crime classification, they are not immune to individual biases and variability. The discrepancies observed between police classifications and Brå's expert annotations suggest that subjective interpretations lead to inconsistent identification of hate crimes.

Model classification issues. Our model achieves a significantly higher precision overall. However, the model does exhibit limitations that contribute to a higher number of false positives. Overclassification likely results from the model's sensitivity to certain language patterns strongly associated with hate crimes in the training data. The model may over-rely on keywords or phrases that frequently appear in hate crime tagged police reports. It does not demonstrate deeper contextual understanding for semantic analysis of some police reports. The model has a low false negative rate. However, there are a few instances concerning under-represented groups or less common hate crime expressions where the model has misclassified. Limited data on certain types of hate crimes can result in poor generalization on police reports that differ from the dominant patterns. While the model's performance in annotating hate crime motives is strong within the subset of flagged cases–surpassing police classification accuracy in that specific context–it is important to acknowledge uncertainties in measurement. Unobserved false negatives, sampling variation, and the inherent difficulties in reliably estimating the F1-score across the entire dataset introduce some caution. Thus, although the model shows promise and often outperforms the police in identified scenarios, its overall superiority must be interpreted with an understanding of these limitations and potential underestimations of its error rate.

4.2 RQ2: Are there systematic biases with respect to crime codes, police region, or category of hate crime motivation?

We examine the distribution of errors across crime code, police region, and hate crime motivation categories for the year 2022.

4.2.1 Crime codes. We examined the precision of the police and model across different crime codes. Crime codes are used across the justice system to facilitate uniform and reliable registration of crimes by authorities and for official statistics [13]. There are 16 crime codes across 6 broader categories with at least 100 occurrences in the dataset (see Appendix A).

Figure 2 illustrates the precision of the police and model in classifying hate crimes for various crime codes. Precision is defined as the proportion of police and model predicted hate crimes that were also confirmed as hate crimes by Brå for each crime code. For example, for police reports marked as hate crimes for the crime related to harassment (9436), the police achieved a precision of 72%; whereas, the model achieved a precision of 95%. Across almost all crime codes, the model precision was considerably higher, with one notable exception. As explained in Section 2.1, agitation against

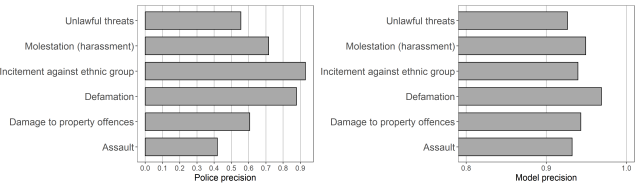


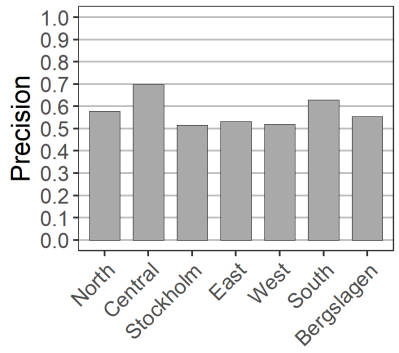
Figure 2: Proportion of cases by crime code identified as hate crimes by police (left) and the model (right), and confirmed by Brå expert annotators. Note that the x-axis for the model begins at 0.8.

a population group (1604) is one of three specific hate crime offenses [82, Ch 16, § 8]. Such offense did not see a clearly improved performance between the police (92.9%) and the model (94.0%) precision; by definition it is a hate crime so classifying it as a hate crime is straightforward for police and model.

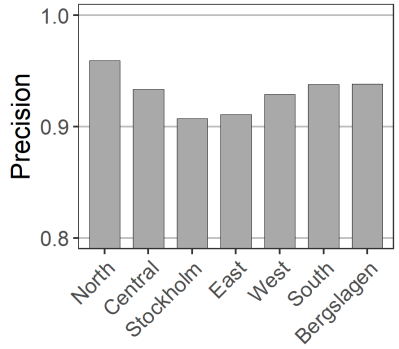
4.2.2 Police regions. In addition to crime codes, we analyzed the distribution of errors made by both the police and the model across different police regions in Sweden. Sweden is divided into seven organizationally different policing regions: Nord [North], Mitt [Central], Stockholm, Öst [East], Väst [West], Syd [South], and Bergslagen. We aim to identify potential differences in the reporting practices and systematic errors in police annotations across these regions. Unlike research in hate speech detection, we define this bias category based on geography and policing structures not for differences in speech, such as varying dialects [74], or social stereotypical language groups [27].

Police precision does vary across region, with the Stockholm region exhibiting a slightly lower precision compared to other regions. Specifically, the police in Stockholm have a higher rate of cases flagged as hate crimes that are not corroborated by Brå's annotations. Model precision was higher overall than police precision, but still had variation across police region. Stockholm equally had lower model precision. While research has shown that the Stockholm police have previously taken an approach that is more restrictive than legal guidance [2, 4], it may be that the substantially larger sample size in Stockholm may contribute to the lower precision, potentially due to a higher diversity of cases or increased complexity in urban settings.

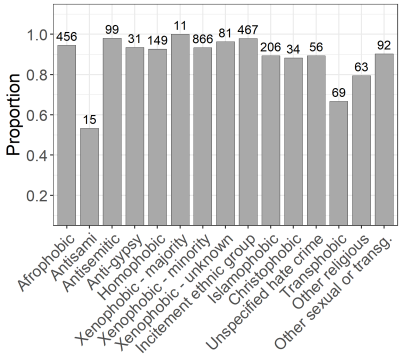
To systematically assess whether crime code and region influence classification outcomes, we employed a logistic regression model (Appendix B). Classification performance varies significantly across regions, with Stockholm police showing generally lower accuracy. A similar but less pronounced effect is observed for the model, indicating that there is likely more difficult cases in the Stockholm region. It is important to note that our findings do not necessarily reflect a *systematic* regional bias but rather may be an artifact of specific case distributions in 2022. One likely explanation is the difficulty of assessing certain incidents that were more prevalent in Stockholm during 2022. A significant number of Quran-burnings and related riots took place in Stockholm that year [51], which we have shown are more challenging for both the police and model to classify (Section 4.1.3).



(a) Police Precision



(b) Model Precision



(c) Model Hate Crime Type Precision

Figure 3: Precision of police (a) and model (b) by police region. (c) Proportion of police reports classified by Brå as hate crimes that the transformer model predicted correctly. Numbers on top of bars indicate number of observations.

4.2.3 Category of hate crime motivation. The category of hate crime motivation is also recorded to distinguish the demographic subject to the incident. We identified the proportion of actual hate crimes that the model accurately classified across several dominant hate crime categories (Figure 3). However, we do not present a similar breakdown for police classifications, primarily because we lack Brå annotations for incidents the police did *not* label as hate crimes. As a result, direct comparisons by hate crime motivation category are not feasible for the police data.

As identified in Section 4.1.4, the model has a higher false negative rate for anti-Sámi hate crimes, likely due to a lack of contextual understanding from minimal training data. Another example shown by this analysis is the lower performance of the model with respect to transphobic hate crimes. Brå has been recording hate crimes against transgender identity or expression since 2008, although there are only 511 confirmed instances in our training data, therefore there are fewer incidents of transphobic hate crimes compared to other categories.

These findings reinforce that classification challenges are not uniform. Certain crime codes are consistently harder to accurately identify as hate crimes for both the police and model. The presence of these difficult categories highlights the importance of context-sensitive interpretation and the need for richer data sources.

4.3 RQ3: What are the potential legal and ethical implications of deploying this model in the criminal justice system?

4.3.1 Distinguishing statistical estimation from policing classification. Before deploying text classification models to address hate crimes in the criminal justice system, it is critical to clearly define the intended purpose and expected benefits [75, 78]. Our findings show that the model often outperforms police classifications in hate crime identification. In such statistical estimation scenarios, where aggregate patterns rather than individual cases drive decision-making, the risk of infringing individual rights is relatively low. The primary ethical and legal considerations revolve around ensuring data protection and transparency of overall metrics rather than risks of individual classifications. Stakeholders may be primarily interested in the overall accuracy, stability, and representativeness of the estimations. The model's superior precision can help Brå allocate resources more effectively, identify under-reported hate crimes, and inform policy deliberation without directly affecting due process for individual suspects or victims. The use of such tools could also help government decision-makers analyze the limitations of Swedish hate crime law and its enforcement, leading to improved policymaking and potential law reform. It can also reduce manual workload for police officers and Brå, and offer a consistent and objective decision-making tool [60].

However, the stakes rise considerably if this approach were adopted for classifying individual hate crime cases in front-line law enforcement. When individual investigative or prosecutorial decisions hinge on whether a crime is classified as hate-motivated, every misclassification can carry real-world consequences. Such deployment should be driven by a demonstrated need to enhance justice and effectiveness, rather than merely seeking efficiency gains [75, 97].

4.3.2 Balancing efficiency with procedural fairness. Although greater accuracy in hate crime classification can significantly improve the allocation of resources and the overall effectiveness of the justice system, even minor classification errors in an operational policing

context may produce disproportionate harm. To illustrate how errors in model classification can undermine these aims, we consider two case studies: false positives and false negatives.

False positives: Burning the Quran. Consider a scenario in which the model incorrectly labels the burning of the Quran as a hate crime, even though the incident does not meet the legal threshold for bias motivation. Such a false positive misclassification risks unjustly affecting the accused by altering both the substantive and procedural dimensions of the case. Substantively, identifying an offense as bias-motivated can influence the severity of sentencing or raise its legal seriousness (Section 2.1). Procedurally, the designation may determine whether the case is handled by specialized hate crime investigators or prosecutors, potentially diverting limited resources away from other cases. If such errors cluster in certain regions or offense types, they further entrench inconsistencies in enforcement and potentially erode public trust in both the technology and broader law enforcement institutions.

False negatives: Hate crimes against Sámi. On the other hand, false negatives, such as failing to identify anti-Sámi hate, highlight the parallel danger of overlooking genuine bias incidents. Missing a true hate crime denies victims the protections, acknowledgment, and targeted investigative efforts they deserve. It also distorts the broader view of bias-motivated offenses, thus impeding informed policy responses and contributing to the under-protection of already marginalized communities. A model that systematically under-detects certain forms of hate crime, whether due to sparse training data or implicit biases, undermines its promise of efficiency by perpetuating inequalities and weakening societal confidence in both technology and law enforcement institutions. High false negative rates in specific hate crime categories may reflect resource gaps or deeper data quality issues, as well as embedded biases in detection and classification processes. Such systematic inaccuracies pose risks of broader inequality in the application of these tools, an especially critical concern given that prosecutors and judges may also fail to identify hate-motivated aggravations [40].

4.3.3 *Aligning with existing legal frameworks.*

Data privacy and security. Using crime reports, particularly those documenting hate incidents, necessarily involves handling sensitive personal data. According to the General Data Protection Regulation (**GDPR**) and Swedish guidelines, personal data pertaining to criminal justice can typically only be processed by official authorities [33, 45]. Brå, Polismyndigheten, and Åklagarmyndigheten function as official authorities, ensuring compliance with legal restrictions. Technical safeguards further mitigate the risk of data leakage. One key strategy we employed is to build on pre-trained encoder models rather than training from scratch [16, 28], limiting the introduction of personal details into model weights. These measures maintain compliance with data protection frameworks and foster trust in law enforcement institutions.

Discrimination. The legal framework on hate crimes includes the criminal law prohibition of unlawful discrimination by certain persons, including business operators or employees, public officials, or organizers of public events [82, Ch 16, § 9]. Discrimination is also prohibited in civil law that applies to employers, organizations, and education institutions, private and certain public service providers [87]. However, not all public activities fall within these prohibitions; in particular, law enforcement authorities fall outside the regulated areas [3, 19]. A 2021 government inquiry proposed amending the Discrimination Act to prohibit discriminatory police measures [3, 19]; however, the proposal has not been enacted. While hate crime classification sits at the intersection of criminal and anti-discrimination law, current legal frameworks do not require non-discrimination in police practices. Despite the absence of a formal legal mandate requiring non-discrimination in police practices, ensuring fairness and legitimacy remains critical, particularly where hate crime classification may directly affect how crimes are investigated, prosecuted, and perceived. If an automated model systematically misclassifies particular communities (e.g., the Sámi) or wrongly flags certain demographic groups at higher rates (e.g., Muslims), it could break principles of non-discrimination. These risks underscore the importance of robust oversight, stakeholder engagement, and ongoing monitoring to prevent the exacerbation of inequalities through algorithmic tools.

Procedural rights. Fundamental procedural justice rights are protected in the Swedish Constitution [83] and guaranteed under the European Convention on Human Rights [26]. Every individual retains protection from unlawful or arbitrary arrest or detention [26, Art 6; 83, Art 8] and a right to a fair trial [26, Art 6; 83, Art 9], underscoring the risks posed by potential misclassifications in automated systems. If a model were to incorrectly flag an individual in a hate crime investigation, it could lead to unwarranted suspicion, thereby affecting the presumption of innocence–a cornerstone of fair trial standards [8, 26]. If these misclassifications were to enter the public domain, the perceived guilt of the accused may be amplified prematurely, further undermining due process before any legal determination is made [8]. Moreover, automation bias can reinforce these errors, as human operators tend to trust machine-generated outputs as more authoritative or objective, thereby reducing the likelihood of meaningful review [20, 81].

EU AI Act. The European Union's AI Act introduces new regulations for AI applications based on their risk profiles [34]. It designates certain AI systems as *high-risk*, requiring heightened obligations and oversight. AI systems used by law enforcement authorities are listed as potential high-risk AI system [34, Annex III(6) pursuant to Art 6(2)]. However, there are some situations in which an AI system listed in Annex III may not be considered high-risk where it "does not pose a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making", including if it is "intended to perform a narrow procedural task" or "intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review" [34, Art 6(3)]. Even if a provider believes their AI system should not be classified as high-risk, they are required to document its assessment and register the system [34, Art 6(4)]. A model for hate crime classifications, in some applications, could meet the definition of a high-risk system. The central concern is whether the system *materially influences* or *replaces* human decision-making

in ways that affect suspects' or victims' fundamental rights. Based on our observations in this paper, efficient and responsible use of such model would not displace or dilute the responsibility of law enforcement and should serve the narrow procedural task of flagging police reports that may be hate crimes.

Transparency and explainability. If the model were to evolve from a statistical estimation tool into a practical mechanism for classifying individual hate crime cases, transparency and explainability concerns become much more pronounced. In an operational context, both the users and those subject to the model classifications should be able to understand how classifications are reached. Under the GDPR, individuals have a right to an explanation if they are subject to automated processing which produces legal or similarly significant effects [33, Art 15, 22; 31, 32]. Even if a human participates in the procedure, it is solely automated processing if they are unable to "influence the causal link between the automated processing and the final decision" [31]. In our view, such models should not be used without meaningful human oversight. While the model may be "black-box" in nature, stakeholders are still likely to have sufficient understanding of a tool's mechanisms so that accountability for decisions remains firmly with human actors and post-hoc analysis can improve interpretability [73, 96, 97]. If well executed, the ability to scrutinize the model and its classifications may lead to more robust police accountability. Transparency and explanations of decisions are also necessary for procedural justice, which requires individuals to understand the reasons behind decisions that affect them in order to ensure that affected individuals are in a position to challenge those decisions [18, 77]. It is particularly important given hate crime victims are often marginalized members of society and both are "more likely both to experience hate-motivated violence and to experience discrimination at the hands of the police" [19, 29]. Accuracy alone will not suffice; all stakeholders should be able to meaningfully understand and scrutinize how decisions are made.

5 Conclusion

This paper evaluates a model for text classification that improves hate crime detection relative to manual police annotations. By reviewing specific misclassification cases–such as incidents involving the Sámi or Quran burnings–we illustrate how data sparsity, contextual nuances, and cultural sensitivities can challenge text classification methods. By systematically analyzing the legal and ethical implications of such a model, we have highlighted the importance of robust oversight, transparency, and accountability. With continued refinement and careful integration of domain expertise, we believe machine learning methods can help build fairer, more consistent processes for identifying complex crimes, such as hate crimes, in the criminal justice system.

Research Ethics and Impact Statement

This research was conducted as part of a partnership between Brå (Swedish National Council for Crime Prevention), which functions as the Swedish Government's body of expertise within the judicial system, and researchers at Uppsala University.

The following aspects were key ethical considerations for this research:

- Data protection and privacy: Stringent measures were implemented to comply with the General Data Protection Regulation (GDPR) to protect participants' privacy and data. As a direct consequence of these requirements, Brå exercises exclusive control over the datasets, as well as the models derived from them, and no component of this information will be publicly disclosed.
- Compliance with Brå policies: The research adhered to all relevant policies issued by Brå, particularly concerning data handling and the responsible use of crime data.
- Swedish Research ethics approval: The study received clearance ensuring it meets national ethical standards for research involving human subjects.

The ethical review was conducted by the Swedish Ethical Review Authority in accordance with Act (2003:460) concerning the ethical review of research involving humans. The Swedish Ethical Review Authority is a government agency independent of the University and assesses whether the research application is in line with Swedish statutes. For this review, we submitted an ethical application containing a general research plan, a CV for the responsible researcher and a standardised form with questions on, for example, research questions and aims, methods used, ethical considerations and a time plan. These documents were then considered by the Swedish Ethical Review Authority. The ethical review was assigned the number 2022-05586-01 and was approved on 25 October 2022.

Acknowledgments

This research was supported by Länsförsäkringar's Research Foundation, Grant P1.22. We are especially grateful to Brå, the Swedish National Council for Crime Prevention, for providing access to data and supporting this collaboration. We thank Louise Ekström, Thomas Hvitfeldt, Jon Lundgren, Aravella Lejonstad, Levent Kemetli and Per Ahlström for their valuable insights and assistance throughout the project. Finally, we warmly acknowledge the peaceful visit to Husarö and its gracious local residents, who each summer patiently welcome an unexpected gathering of international PhD students.

References

[1] Sophia Adams-Bhatti and Holli Sargeant. 2024. Algorithms in the Justice System: Current Practices, Legal and Ethical Challenges. In *The Law of Artificial Intelligence* (2 ed.), Matt Hervey and Matthew Lavy (Eds.). Sweet & Maxwell, London.

[2] Åklagarmyndigheten. 2022. Hatbrott [Hate crimes]. Rättslig Vägledning 2022:11 [Legal Guidance].

[3] Arbetsmarknadsdepartementet. 2021. Slutbetänkande av Utredningen ett utökat skydd mot diskriminering [Final Report of the Inquiry into extended protection against discrimination]. Sou 2021:94.

[4] Kıvanç Atak. 2020. 'Inappropriate but not crime'? Policing racial hatred in Sweden. *Nordic Journal of Criminology* 21, 1 (2020), 32–48.

[5] Jeannine Bell. 2002. *Policing Hatred: Law Enforcement, Civil Rights, and Hate Crime.* New York University Press, New York.

[6] Erik Berggren and Anders Neergaard. 2015. *Populism: Protest, Democratic Challenge and Right-Wing Extremism.* Routledge, London, 187–217.

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (2018), 3–44.

[8] Kelly Blount. 2021. Applying the Presumption of Innocence to Policing with AI. *International Review of Penal Law* 92, 1 (2021), 33–48.

[9] Aaron Boxerman and Isabella Kwai. 2023. What's Happening With the Quran Burnings in Sweden. https://www.nytimes.com/article/sweden-denmark-quran-burnings.html

[10] Brå. 2009. Hatbrott 2008: Polisanmälningar där det i motivbilden ingår etnisk bakgrund, religiös tro, sexuell läggning eller könsöverskridande identitet eller uttryck [Hate Crimes 2008: Police reports where the motive includes ethnic background, religious belief, sexual orientation or transgender identity or expression] (Rapport 2009:10).

[11] Brå. 2021. Polisanmälda hatbrott 2020: En sammanställning av de ärenden som hatbrottsmarkerats av polisen [Police-reported hate crimes 2020: A compilation of cases marked as hate crimes by the police] (Rapport 2021:17).

[12] Brå. 2024. Hatbrott mot samer [Hate crimes against the Sámi people] (Rapport 2024:5).

[13] Brottsförebyggande rådet. 2024. Klassificering av brott: Anvisningar och regler [Classification of Crimes: Instructions and Rules]. v. 12.1.

[14] David Buil-Gil, Angelo Moretti, and Samuel H. Langton. 2022. The accuracy of crime statistics: assessing the impact of police data bias on geographic crime analysis. *Journal of Experimental Criminology* 18, 3 (2022), 515–541.

[15] Bryan D Byers, Kiesha Warren-Gordon, and James A Jones. 2012. Predictors of hate crime prosecutions: An analysis of data from the national prosecutors survey and state-level bias crime laws. *Race and Justice* 2, 3 (2012), 203–219.

[16] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21).* USENIX Association, Online, 2633–2650. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[17] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163.

[18] Danielle Citron. 2008. Technological Due Process. *Washington University Law Review* 85, 6 (2008), 1249–1313.

[19] Civil Rights Defenders. 2024. *Joint submission to the UN Universal Periodic Review of Sweden: 49th Session of the UPR Working Group of the Human Rights Council April-May 2025.* Civil Rights Defenders, Sweden, Stockholm, Sweden.

[20] Jennifer Cobbe. 2018. Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making. *Legal Studies* 39, 4 (2018), 636–655.

[21] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Acm, Halifax, Canada, 797–806.

[22] Hannah Couchman. 2019. *Policing By Machine: Predictive Policing and the Threat to Our Rights.* Technical Report. Liberty. https://www.libertyhumanrights.org.uk/issue/policing-by-machine/

[23] United Nations Human Rights Council. 2010. Report of the Working Group on the Universal Periodic Review Sweden. A/HRC/15/11, https://documents.un.org/doc/undoc/gen/g10/144/53/pdf/g1014453.pdf.

[24] United Nations Human Rights Council. 2015. Report of the Working Group on the Universal Periodic Review Sweden. A/HRC/29/13, https://documents.un.org/doc/undoc/gen/g15/076/76/pdf/g1507676.pdf.

[25] United Nations Human Rights Council. 2020. Report of the Working Group on the Universal Periodic Review Sweden. A/HRC/44/12, https://documents.un.org/doc/undoc/gen/g20/069/40/pdf/g2006940.pdf.

[26] Council of Europe. 1950. Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights). ETS No. 005.

[27] Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics* 11 (2023), 300–319. doi:10.1162/tacl_a_00550

[28] Sunny Duan, Mikail Khona, Abhiram Iyer, Rylan Schaeffer, and Ila R Fiete. 2024. Uncovering Latent Memories: Assessing Data Leakage and Memorization Patterns in Frontier AI Models. arXiv:2406.14549 [cs.CV]

[29] Caroline Erentzen and Regina Schuller. 2020. Exploring the Dark Figure of Hate: Experiences with Police Bias and the Under-reporting of Hate Crime. *Canadian Journal of Criminology and Criminal Justice* 62, 2 (2020), 64–97.

[30] Virginia Eubanks. 2018. *Automating Inequality.* St. Martin's Press, New York.

[31] European Court of Justice. 2023. Opinions of AG Pikamäe, SCHUFA Holding and Others II (C-634/21). Ecr 000.

[32] European Court of Justice. 2024. Opinion of AG de la Tour, Dun & Bradstreet Austria (C-203/22). Ecr 000.

[33] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119/1.

[34] European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). OJ L 2024/1689, 1.

[35] Andrew Guthrie Ferguson. 2016. Policing Predictive Policing. *Washington University Law Review* 94, 5 (2016), 1109–1190.

[36] Phebe Fjellström. 1985. *Samernas Samhälle.* Norsteds, Sweden.

[37] Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.* 51, 4, Article 85 (July 2018), 30 pages. doi:10.1145/3232676

[38] Görel Granström and Karin Åström. 2017. *Lifecycle of a Hate Crime Country Report for Sweden.* Technical Report. Umeå University.

[39] Mika Hagerlid, Görel Granström, and Louise Gustavsson. 2024. Success Factors for Hate Crime Investigation in Sweden. In *Dynamics of Hate: Examining Interdisciplinary Perspectives.* International Network for Hate Studies Biennial Conference, Cape Town, 47–47.

[40] Mika Hagerlid and Görel Granström. 2023. Hate Crime Investigation and Sentencing in Sweden: What Have We Learned in the Past 20 Years? *European Journal on Criminal Policy and Research* (2023), 18 pages.

[41] Stevie-Jade Hardy, Neil Chakraborti, and Ilda Cuko. 2020. More Than A Tick-box? The Role Of Training In Improving Police Responses To Hate Crime. *British Journal of Community Justice* 16, 1 (2020), 4–20.

[42] House of Commons Public Administration Select Committee. 2014. Caught red-handed: Why we can't count on Police Recorded Crime statistics. Thirteenth Report of Session 2013–14.

[43] Polismyndigheten i Stockholms län. 2004. Kartläggning av hatbrott. Rädsla för det främmande [A Study on Hate Crimes. Fear of the Unknown]. Länskriminalpolisen: integrationssektionen.

[44] Polismyndigheten i Stockholms län. 2007. Hatbrott i City polismästardistrikt 13 februari till 9 maj 2007 [Hate crimes in City Police District between 13th and 9th May 2007]. LKP-KUT no. 66/07..

[45] Imy. 2021. PRättsligt ställningstagande – innebörden av begreppet "personuppgifter som rör lagöverträdelser som innefattar brott" i artikel 10 i dataskyddsförordningen [Legal position - the meaning of the term "personal data relating to offenses involving crimes" in Article 10 of the Data Protection Regulation]. Imyrs 2021:1.

[46] James B Jacobs and Kimberly Potter. 2000. *Hate Crimes: Criminal Law and Identity Politics.* Oxford University Press, Oxford.

[47] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference,*, Vol. 67. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 43:1–43:23.

[48] Klara Klingspor. 2008. The Challenges of Collecting Statistical Data in the Field of Hate Crime: The Case of Sweden. In *Hate Crime: Papers from the 2006 and 2007 Stockholm Criminology Symposiums*, Jo Goodey and Kauko Aromaa (Eds.). European Institute for Crime Prevention and Control, affiliated with the United Nations, Helsinki, Finland, 40–55.

[49] Brendan Lantz, Andrew S Gladfelter, and R Barry Ruback. 2019. Stereotypical hate crimes and criminal justice processing: A multi-dataset comparison of bias crime arrest patterns by offender and victim race. *Justice Quarterly* 36, 2 (2019), 193–224.

[50] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[51] Göran Larsson and Christer Mattsson. 2024. Public Opinions on Freedom of Speech and Prohibited Hate Speech against Islam and Muslims: Rasmus Paludan,

Burning of the Quran and Swedish Media. *Temenos - Nordic Journal for the Study of Religion* 60, 1 (June 2024), 131–156. doi:10.33356/temenos.136834

[52] Clara S Lewis. 2014. *Tough on Hate?: The Cultural Politics of Hate Crimes.* Rutgers University Press, New Brunswick.

[53] Kristian Lum and William Isaac. 2016. To Predict and Serve? *Significance* 13, 5 (Oct. 2016), 14–19.

[54] Jon Lundgren and Aravella Lejonstad. 2023. *Polisanmälda hatbrott 2022 - en sammanställning av de ärenden som hatbrottsmarkerats av polisen [Hate crimes in police reports 2022 - a summary of reports marked as hate crimes by the police].* Technical Report. Brottsförebyggande rådet (BRÅ). Report nr 2023:16.

[55] Christopher Lyons and Aki Roberts. 2014. The difference "hate" makes in clearing crime: An event history analysis of incident factors. *Journal of Contemporary Criminal Justice* 30, 3 (2014), 268–289.

[56] Sean MacAvaney, Hong Yao, Enyan Yang, Kyle Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one* 14, 8 (2019), e0221152. doi:10.1371/journal.pone.0221152

[57] Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. arXiv:2007.01658 [cs.CL]

[58] Sandra Mayson. 2019. Bias In, Bias Out. *Yale Law Journal* 128, 8 (2019), 2122–2473.

[59] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163.

[60] Aida Mostafazadeh Davani, Leigh Yeh, Mohammad Atari, Brendan Kennedy, Gwenyth Portillo Wightman, Elaine Gonzalez, Natalie Delong, Rhea Bhatia, Arineh Mirinjian, Xiang Ren, and Morteza Dehghani. 2019. Reporting the Unreported: Event Extraction for Analyzing the Local Representation of Hate Crimes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Association for Computational Linguistics, Hong Kong, China, 5753–5757.

[61] Nordiska Museet. 2024. *Norbor: Liv och rörelse under 500 år [Northerners: Life and movement over 500 years].* Nordiska Museet i samarbete med Makadam förlag, Stockholm, Sweden.

[62] James J. Nolan, Stephen M. Haas, and Jessica S. Napier. 2011. Estimating the Impact of Classification Error on the "Statistical Accuracy" of Uniform Crime Reports. *Journal of Quantitative Criminology* 27, 4 (2011), 497–519.

[63] James J. Nolan III, Jack McDevitt, Shea Cronin, and Amy Farrell. 2004. Learning to See Hate Crimes: A Framework for Understanding and Clarifying Ambiguities in Bias Crime Classification. *Criminal Justice Studies* 17, 1 (2004), 91–105.

[64] Organization for Security and Co-operation in Europe. 2022. *Hate Crime Laws: A Practical Guide* (2 ed.). Osce/odihr, Warsaw.

[65] Barbara Perry and Shahid Alvi. 2012. 'We are all vulnerable': The *in terrorem* effects of hate crimes. *International Review of Victimology* 18, 1 (2012), 57–71.

[66] Frank S. Pezzella and Matthew D. Fetzer. 2017. The Likelihood of Injury Among Bias Crimes: An Analysis of General and Specific Bias Types. *Journal of Interpersonal Violence* 32, 5 (2017), 703—-729.

[67] Anchal Rawat, Santosh Kumar, and Surender Singh Samant. 2024. Hate speech detection in social media: Techniques, recent trends, and future challenges. *WIREs Computational Statistics* 16, 2 (March 2024), e1648. doi:10.1002/wics.1648

[68] Regeringskansliet. 2023. Action programme to combat racism against Sami. https://www.government.se/information-material/2023/02/action-programme-to-combat-racism-against-sami/

[69] Reuters. 2023. Sweden is considering law change to stop public Koran burnings, Aftonbladet reports. https://www.reuters.com/world/europe/sweden-is-considering-making-koran-burnings-illegal-aftonbladet-2023-07-06/

[70] Rashida Richardson, Jason M. Schultz, and Kate Crawford. 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. *New York University Law Review Online* 94 (2019), 15–55.

[71] Rikspolisstyrelsen and Åklagarmyndigheten. 2008. Redovisning av uppdrag angående hatbrott [An account of a commission concerning hate crime]. Dnr: POA-426-547/07.

[72] Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat (Eds.). Association for Computational Linguistics, Seattle, Washington (Hybrid), 154–169. doi:10.18653/v1/2022.woah-1.15

[73] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[74] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 1668–1678. doi:10.18653/v1/P19-1163

[75] Teresa Scantamburlo, Andrew Charlesworth, and Nello Cristianini. 2019. *Machine Decisions and Human Consequences.* Oxford University Press, Oxford, 49–81.

[76] Mårten Schultz. 2023. Understanding why burning the Qur'an isn't illegal in Sweden means looking at the country's long-held commitment to freedom of expression. http://theconversation.com/understanding-why-burning-the-quran-isnt-illegal-in-sweden-means-looking-at-the-countrys-long-held-commitment-to-freedom-of-expression-211689

[77] Andrew Selbst and Solon Barocas. 2018. The Intuitive Appeal of Explainable Machines. *Fordham Law Review* 87, 3 (2018), 1085.

[78] Esther Shein. 2018. The dangers of automating social programs. *Commun. ACM* 61, 10 (Sept. 2018), 17–19.

[79] Keller G. Sheppard, Nathaniel L. Lawshe, and Jack McDevitt. 2021. Hate Crimes in a Cross-Cultural Context.

[80] Jawaid Ahmed Siddiqui, Siti Sophiayati Yuhaniz, and Zulfiqar Ali Memon. 2024. A Comparative Study of Automatic Hate Speech Detection Using Machine Learning. In *2024 IEEE 1st Karachi Section Humanitarian Technology Conference (KHI-HTC).* IEEE, New York, USA, 1–7. doi:10.1109/KHI-HTC60760.2024.10482049

[81] Linda Skitka, Kathleen Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51, 5 (1999), 991–1006.

[82] Sveriges Riksdagen. 1965. Brottsbalken [Criminal Code]. SFS nr 1962:700 (Official English Translation).

[83] Sveriges Riksdagen. 1974. Kungörelse om beslutad ny regeringsform [Constitution of Sweden: Instrument of Government]. SFS nr 1974:152 (Official English Translation).

[84] Sveriges Riksdagen. 1991. Yttrandefrihetsgrundlag [Constitution of Sweden: Fundamental Law on Freedom of Expression]. SFS nr 1991:1469 (Official English Translation).

[85] Sveriges Riksdagen. 1993. Regeringens proposition: Åtgärder mot rasistisk brottslighet och etnisk diskriminering i arbetslivet [Government bill: Measures against racist crime and ethnic discrimination in working life]. 1993/94:101.

[86] Sveriges Riksdagen. 2002. Regeringens proposition: Hets mot folkgrupp, m.m. [Government bill: Incitement against an ethnic group, etc.]. 2001/02:59.

[87] Sveriges Riksdagen. 2008. Diskrimineringslag [Discrimination Act]. SFS nr 2008:567 (Non-Official English Translation).

[88] Sveriges Riksdagen. 2018. Regeringens proposition: Ett utvidgat straffrättsligt skydd för transpersoner [Government bill: Expanded criminal law protection for transgender people]. 2017/18:59.

[89] Sveriges Riksdagen. 2021. Lag om den officiella statistiken [Official Statistics Act]. SFS nr 2001:99.

[90] Armani Syed. 2023. Why Quran Burning Is Making Sweden and Denmark So Anxious. https://time.com/6303348/quran-burning-sweden-denmark/

[91] Sámetinget. 2024. Hatbrott mot samer [The Sámi in Sweden]]. https://www.sametinget.se/samer

[92] UK Statistics Authority. 2014. Assessment of compliance with the code of practice for official statistics. Statistics on crime in England and Wales. Assessment Report 268.

[93] Hannes Waldetoft, Jakob Torgander, and Måns Magnusson. 2025. Prediction-powered estimators for finite population statistics in highly imbalanced textual data: Public hate crime estimation. arXiv:2505.04643 [cs.CL] https://arxiv.org/abs/2505.04643

[94] Scott M Walfield, Kelly M Socia, and Rachael A Powers. 2017. Religious motivated hate crimes: Reporting to law enforcement and case outcomes. *American Journal of Criminal Justice* 42 (2017), 148–169.

[95] Mark Austin Walters. 2014. Conceptualizing 'hostility'for hate crime law: Minding 'the Minutiae'when interpreting Section 28 (1)(a) of the Crime and Disorder Act 1998. *Oxford Journal of Legal Studies* 34, 1 (2014), 47–74.

[96] Adrian Weller. 2019. Transparency: Motivations and Challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (Eds.). Springer, Cham, 23–40.

[97] Miri Zilka, Holli Sargeant, and Adrian Weller. 2022. Transparency, Governance and Regulation of Algorithmic Tools Deployed in the Criminal Justice System: a UK Case Study. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(Aies '22).* Acm, New York, NY, USA, 880–889.

[98] Marta Ziosi and Dasha Pruss. 2024. Evidence of What, for Whom? The Socially Contested Role of Algorithmic Bias in a Predictive Policing Tool. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency.* Acm, Rio de Janeiro, Brazil, 1596–1608.

A Crime codes with at least 100 occurrences of police reports tagged as hate crimes in 2022

There are 16 crime codes that appeared in at least 100 police reports tagged as hate crimes in 2022. These can be grouped into 6 types of crimes under the Swedish Criminal Code: assault, unlawful threats, molestation (harassment), defamation, damage to property offenses, offenses against public order [82].

Count	Code	Swedish Description	English Translation
708	1604	Hets mot folkgrupp	Agitation against a population group
561	1209	Skadegörelse, övrigt klotter	Damage to property, other graffiti
100	9458	Ofredande, mot kvinna 18 år eller äldre, är eller har varit bekanta genom annan slags relation	Molestation (Harassment) of a woman over 18 years of age, are or have been acquainted through another kind of relationship
314	9459	Ofredande, mot kvinna 18 år eller äldre, obekanta	Molestation (Harassment) of a woman over 18 years of age, stranger
284	9463	Ofredande, mot man 18 år eller äldre, obekanta	Molestation (Harassment) of a man over 18 years of age, stranger
246	1212	Annan skadegörelse (ej klotter)	Other damage to property (not graffiti)
228	513	Ärekränkningsbrott; förtal, förolämpning, förtal av avliden, mot man 18 år eller äldre, ej internetrelaterat	Defamation, insulting behaviour, defamation of a deceased person, against a man over 18 years of age, not internet-related
200	9447	Olaga hot, ej internetrelaterat, mot man 18 år eller äldre, obekanta	Unlawful threat, not internet related, against man over 18 years of age, stranger
145	1205	Skadegörelse, mot stat, kommun, landsting, ej klotter	Damage to property, against state, municipality, regional council, not graffiti
168	414	Ofredande mot grupp	Molestation (Harassment) of a group
153	357	Misshandel, annan än grov, mot man 18 år eller äldre, obekanta, utomhus	Assault, other than aggravated, against a man over 18 years of age, stranger, outdoors
139	511	Ärekränkningsbrott; förtal, förolämpning, förtal av avliden, mot kvinna 18 år eller äldre, ej internetrelaterat	Defamation, insulting behaviour, defamation of a deceased person, against a woman over 18 years of age, not internet-related
133	1201	Skadegörelse, på motorfordon, ej genom brand	Damage to property, to motor vehicles, not by fire
103	428	Ofredande mot flicka under 18 år	Molestation (Harassment) of a girl under 18 years of age
111	429	Ofredande mot pojke under 18 år	Molestation (Harassment) of a boy under 18 years of age
107	9443	Olaga hot, ej internetrelaterat, mot kvinna 18 år eller äldre, obekanta	Unlawful threat, not internet related, against woman over 18 years of age, stranger

Table 2: Crime codes with at least 100 occurrences within the reports marked by the police as hate crimes year 2022.

B Logistic regression results

To systematically compare the performance of the police and the transformer model, we fit a logistic regression in which the response variable was *correct classification* versus *incorrect classification*, as determined by Brå's 2022 annotations (Table 3). We included crime codes and police region as explanatory variables. In order to maintain a sufficiently aggregated level of analysis, crime codes (four digits) were truncated to their first two digits, preserving the high-level category (e.g., "assault") while removing granular information about age or gender of the victim. We then fit two separate logistic models: one for the police classifications and one for the transformer model predictions. For the transformer model, we restricted the analysis to cases where the model predicted "hate crime," so that the model's probability estimates aligned with Brå's true/false labels in a comparable subset of reports.

A likelihood ratio test was used to assess the overall significance of *police region* within both logistic models, revealing statistically significant differences in performance across regions for both the police ($p_{police} = 2.5 \times 10^{-13}$) and the transformer model ($p_{model} = 0.00018$). After further inspection, the least square means for *police region* indicated the lowest performance in the Stockholm region for both classifications (Figure 4). Hence, we made a pairwise comparison between Stockholm and the other regions (Table 4), for which the results indicated a more pronounced difference for the police than the transformer model.

	Police		Transformer model	
Variable	Coefficient	SE	Coefficient	SE
Intercept	0.109	0.082	2.40	0.206
Crime Code 1604	2.10	0.160	-0.0742	0.232
Crime Code 0357	-0.830	0.180	-0.160	0.547
Crime Codes 0411+0429	0.563	0.134	0.0273	0.315
Crime Codes 0511+0513	1.630	0.180	0.907	0.397
Crime Codes 9458+9459+9443+9447+9463	0.179	0.0923	0.051	0.242
North	0.649	0.167	1.041	0.440
Central	0.934	0.182	0.333	0.332
East	0.426	0.138	0.202	0.279
West	0.312	0.116	0.629	0.282
South	0.569	0.111	0.546	0.238
Bergslagen	0.371	0.154	1.08	0.440

Table 3: Results from logistic regressions. Crime codes grouped by first two digits. Crime codes starting with 12 and Stockholm police regsion are reference levels.

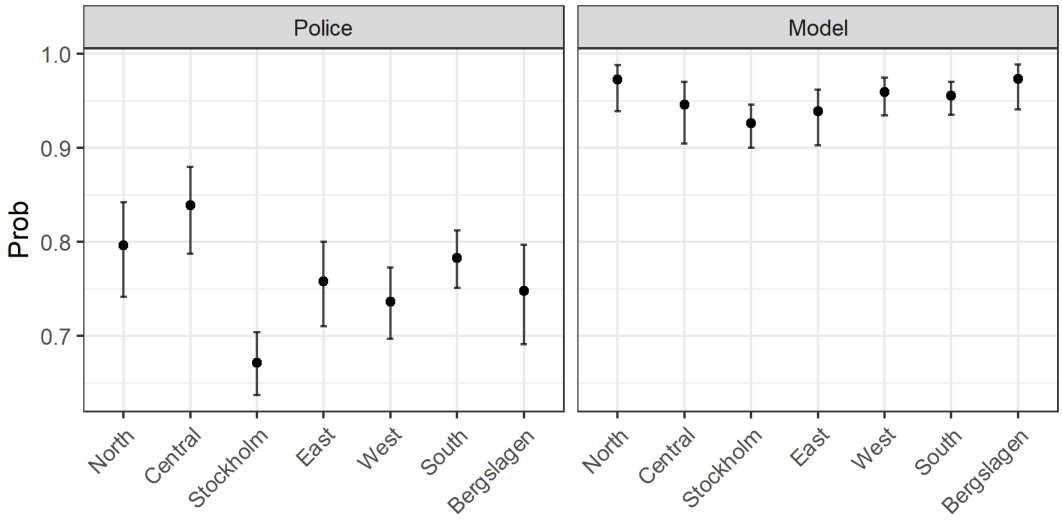


Figure 4: Least square means for Police region, averaged over Crime code, on response scale (probabilities). Error bars represent asymptotic confidence intervals, on 95%-level.

	Police		Transformer model	
Contrast	Odds ratio	p-value	Odd ratio	p-value
North/Stockholm	1.9	0.0006	2.8	0.087
Central/Stockholm	2.5	<0.0001	1.4	0.77
East/Stockholm	1.5	0.011	1.2	0.90
West/Stockholm	1.4	0.0036	1.9	0.12
South/Stockholm	2.8	<0.0001	1.7	0.10
Bergslagen/Stockholm	1.5	0.079	2.4	0.072

Table 4: Dunnett multiple comparison of Stockholm versus the other police regions.