Journal of Theoretical and Applied Information Technology
31st March 2025. Vol.103. No.6

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

USING AI TO COUNTER CRIMINAL OFFENCES AGAINST THE WILL, HONOUR, AND DIGNITY OF A PERSON

VIACHESLAV SHEVCHENKO[1], OLEH ZVONAROV[2], LEONID SHCHERBYNA[3], YANA LUTSENKO[4], VOLODYMYR PYVOVAROV[5]

[1]Interregional Academy of Personnel Management, Ukraine

[2]Bila Tserkva National Agrarian University, Chair Public and Legal Disciplines, Faculty of Social Sciences and Humanities, Ukraine

[3]Scientific and Organizational Center National Academy of the Security Service of Ukraine, Ukraine
[4]National Academy of the Security Service of Ukraine, Special Department № 8 Educational and Scientific Institute of State Security, Ukraine

[5]Academician Stashis Scientific Research Institute for the Study of Crime Problems National Academy of Legal Sciences of Ukraine, The Department of Criminal Executive Law Research, Ukraine

E-mail: [1]tymoshyu@gmail.com, [2]zv.helga@ukr.net, [3]l.shcherbina@ukr.net, [4]coolyana@gmail.com, [5]pyvovarov.volodymyr1@gmail.com

ABSTRACT

In the era of the digitalization of society, crimes against the will, honour, and dignity of a person, in particular cyberbullying, blackmail, and threats, are becoming increasingly common. These phenomena are becoming increasingly common due to the anonymity of online communication and the rapid dissemination of information. This creates new challenges for law enforcement agencies, requiring the implementation of innovative technologies, in particular artificial intelligence (AI), for their detection, prevention, and investigation. Artificial intelligence (AI) opens up opportunities for automated analysis of threatening messages and prediction of risks of aggressive behaviour.

The research focuses on assessing the potential of AI in combating such crimes, analysing its effectiveness, and identifying the limitations of technology use in law enforcement practice. The study evaluates the effectiveness of AI algorithms for detecting cyber threats and identifies their key limitations. The aim of the study is to assess the effectiveness of AI algorithms for automated detection of threatening messages, prediction of risks of criminal behaviour compared to traditional methods.

The study employed the following methods: text data analysis using natural language processing (NLP), behavioural pattern modelling to predict risks, and surveys of human rights defenders and lawyers to study their attitudes towards AI in combating crime. The methods used include text data analysis, behavioural pattern modelling, and a sociological survey of human rights activists regarding their attitudes towards AI.

The results showed that AI algorithms demonstrate high accuracy rates in detecting cyber threats, outperforming traditional methods in terms of speed and scalability. NLP algorithms achieved 85% accuracy compared to 75% in manual moderator analysis, confirming their effectiveness. At the same time, a survey of specialists revealed a number of ethical and legal limitations in the implementation of these technologies, in particular, 60% of respondents indicated the need for strict regulation of AI, and 35% emphasized the risk of false accusations.

The academic novelty of the study is the interdisciplinary approach that integrates technological analysis with legal aspects to assess the effectiveness of AI, but also to identify obstacles to its practical application. The novelty of the study lies in combining the analysis of the effectiveness of AI with an assessment of the possibilities of its actual use in the fight against cyber threats. Further research should focus on adapting algorithms to the changing conditions of the digital environment and developing regulatory mechanisms for their implementation in compliance with human rights. Further research should improve AI algorithms and reduce the false positive rate.

Keywords: *Artificial Intelligence, Criminal Offences, Cyber Threats, Legal Ethics, Text Analysis.*

1. INTRODUCTION

The modern digital era brings both significant benefits and serious challenges. Crimes against the will, honour, and dignity of a person, such as cyberbullying, blackmail and threats, are becoming more widespread because of the anonymity of the network. This requires new approaches to protecting society. Previous studies [2,3] have focused on the effectiveness of AI in detecting threatening messages but have rarely considered them in conjunction with behavioural models, creating a research gap regarding their integration into criminological analysis [4].

AI opens up unique opportunities to counter such threats through the analysis of large data volumes and risk prediction. For example, NLP algorithms detect signs of threats in texts, and risk prediction models estimate the likelihood of criminal acts [1]. At the same time, modern research [5] emphasizes that the effectiveness of such algorithms depends not only on their accuracy but also on the social and legal context, which requires further study.

At the same time, there are some challenges related to data privacy, ethical application, and technical limitations of AI. The lack of a clear regulatory framework also complicates the implementation of these technologies. Since legal mechanisms do not always keep up with the dynamics of cybercrime, automated methods can become key in the preventive control of digital threats [6].

Research into the effectiveness of AI in combating crimes against the will, honour, and dignity of a person is important for improving mechanisms for protecting human rights, analysing cyber threats and developing modern approaches to their neutralization. In this context, it is essential not only to deter cybercrime dynamics existing algorithms but also to analyse their practical application in law enforcement activities and their potential impact on digital security policy.

The aim of the study is to determine the effectiveness of using AI in detecting, predicting and preventing crimes against the will, honour, and dignity of a person, as well as to analyse the legal and ethical barriers that complicate its implementation.

Empirical objectives:

1. Analyse the effectiveness of NLP algorithms in detecting threatening messages based on social media data.
2. Assess the accuracy of models for predicting the risks of criminal behaviour based on behavioural data.
3. Study the attitude of law specialists and law enforcement agencies towards the use of AI to combat crimes against the will, honour, and dignity of a person.

2. LITERATURE REVIEW

AI is increasingly being used to detect and combat crime, entailing debate among scholars about its effectiveness, ethics, and legal aspects. The introduction of these technologies has generated both approval and concern, leading to a variety of approaches to their research.

Current research shows that the effectiveness of using artificial intelligence to detect crimes depends on the specifics of each country's legal system. Some authors point out that the regulation of algorithmic justice has not yet been formed at the proper level, which may cause disputes about its legitimacy.

Rimo [2] analyses the trend towards the increasing criminalisation of prior acts in Spanish criminal law, highlighting that early intervention in crimes thanks to AI can contribute to the prevention of offences. However, the author also notes the risks associated with restricting an individual's rights based on suspicion through algorithmic analysis.

This issue is relevant for the Spanish legal system and other countries that implement preventive justice systems. Automated detection of suspicious actions can lead to undue restrictions on human rights, raising questions about the balance between crime prevention and respect for the principle of the presumption of innocence.

Jin and Salehi [3] support this point by drawing attention to the difficulties faced by public defenders trying to review AI decisions during legal procedure. They note that the lack of transparency in the operation of algorithms makes their legal assessment difficult.

In addition, the lack of access to the AI decision-making mechanism creates significant risks for legal proceedings. Human rights activists and lawyers emphasize the need to increase algorithm transparency to avoid situations where AI makes decisions that cannot be challenged due to the lack of access to its logic.

A study by [4] focuses on public perceptions of the use of AI, particularly in healthcare. Although this area does not directly belong to the domain of criminal law, their findings on the importance of transparency and trust in AI

are relevant to the broader context. Meanwhile, Kieslich and Lünich [5] examine public opinion on the regulation of AI for biometric identification. They point to a high demand for audit and registration of databases, which could form the basis for regulation in criminal law.

In the context of criminal law, these studies show that the implementation of AI in the security sector must consider its technical capabilities and the level of public trust. If citizens doubt the reliability and impartiality of algorithms, their use may provoke public opposition.

The researchers in [6] examine the overall impact of AI on crime, noting that its use opens up new opportunities for crime detection but also creates risks such as cybercrime or algorithmic discrimination. Their findings are partly in line with those of Hardy et al. [7], who deal with the use of AI for suicide prevention, which confirms the broader potential of the technology in risk prediction.

Research suggests that AI can be an effective tool for detecting threats and predicting potential crimes. However, a critical issue is discrimination, where algorithms can show bias towards certain social groups, as evidenced by examples of algorithmic racism or social bias in justice systems.

According to [8] focuses on the unreliability of AI in risk assessment systems, indicating that algorithms can lead to discriminatory decisions, especially against marginalized groups. This is consistent with the findings of Wang et al. [9], who examine algorithmic discrimination in the United States, emphasizing the need for its regulation to reduce bias.

These findings indicate the need to develop mechanisms to control algorithmic decisions that avoid automated discrimination. The research also highlights the role of independent auditing and legal norms that can ensure the fair use of AI.

Kopotun et al. [10] raise the issue of the possibility of perceiving AI as an agent of crime, using the example of American criminal law. This raises an interesting debate about liability in the event of criminal acts committed with the help of algorithms. At the same time, Adam et al. [11] emphasize the potential of AI in creating an open justice system, emphasizing the importance of transparency to ensure citizen trust.

It is worth considering the liability issue in crimes committed using AI. Some authors suggest considering algorithms as a tool and a possible liability subject, which opens up new legal discussions.

The study by [12] examines the impact of genetic factors in the context of medical research, which indicates the importance of an interdisciplinary approach to data analysis. Although this issue is mainly related to the medical field, the methods used to identify links between genetic data and risks can find application in criminal analysis, especially in risk prediction.

In this context, criminal analysis can use similar approaches to assess the risks of criminal behaviour based on the study of social network users' behavioural patterns.

The researchers in [13] study the use of information technology (IT) to improve crime prevention mechanisms in the border regions of southern Ukraine. The authors note that the implementation of innovative approaches, in particular the use of digital platforms for data collection and analysis, contributes to a prompt response to potential threats. This idea has something in common with the use of AI for text and behavioural analysis in our research. However, unlike our focus on threats in the digital environment, their research is more focused on physical security in border regions.

The application of AI in the context of information security can also contribute to increasing the effectiveness of preventing crimes in the digital space. However, the effectiveness of such systems largely depends on the level of cooperation between public and private structures and on regulatory support.

Kortukova et al. [14] analyse the features of the legal regulation of temporary protection in the European Union (EU) in the context of Russian aggression against Ukraine. The study emphasizes the importance of ensuring human rights and the use of legal mechanisms in crisis situations. Although this study focuses on regulatory and legal aspects, it emphasizes the relevance of integrating innovative technologies for effective crisis management, which can also be considered as a complement to our analysis of the ethics and regulatory regulation of AI.

These studies point to the importance of combining legal mechanisms and technological solutions to ensure adequate protection of human rights. Integrating AI into human rights protection and law enforcement should be accompanied by clear ethical standards to avoid abuses and maintain citizens' trust in the justice system.

3. METHODS AND MATERIALS

The study was conducted in three stages to analyse the effectiveness of AI in combating crimes against freedom, honour, and dignity of a person. The first stage included a theoretical analysis of modern approaches to the AI use, in particular, NLP algorithms and behavioural pattern models for risk prediction.

At this stage, alternative approaches to text data analysis and risk prediction were also considered. Other NLP algorithms were considered, including BERT and GPT, but due to computational resource requirements, spaCy and NLTK were chosen. Several machine learning algorithms were analyzed for behavioural pattern modelling. Decision Trees, Random Forests, and SVM were selected due to their high interpretability and efficiency in working with small samples.

The second stage included data collection and systematization: texts of threatening messages from social networks (Facebook, Instagram, Reddit, Twitter), behavioural data for risk modelling, and responses from 20 surveyed human rights defenders.

Data collection was conducted exclusively from open sources, which meets ethical research standards. All text messages were obtained from publicly available posts and discussions, and only profiles whose owners had permitted their activity to be analyzed as part of community monitoring initiatives were used to collect behavioural data.

The third stage consisted of quantitative and qualitative analysis. NLP algorithms analysed texts, machine learning (ML) models assessed behavioural risks, and survey results were used to determine human rights defenders' attitudes toward AI.

It is important to note that the analysis methods used have certain limitations. In particular, NLP algorithms may demonstrate reduced accuracy in cases of hidden threats or non-standard language constructs, and machine learning is limited by the source data quality and possible sampling biases. In addition, the sociological survey was conducted among 20 respondents, which is a relatively small sample but allows for expert assessments of representatives of the human rights community.

3.1. Methods

The study employed the following methods:

- text data analysis using NLP – NLP algorithms helped to automatically detect keywords and phrases indicating threats, blackmail or cyberbullying, as well as classify texts by threat level (low, medium, high);

- behavioural pattern modelling – the risks of criminal behaviour by analysing trends in user behaviour on social networks were predicted using ML algorithms;

- sociological survey of human rights defenders – an online survey of 20 human rights defenders was conducted via Google Forms for collecting data on their attitude towards the use of AI, ethical aspects, as well as advantages and disadvantages of such technologies.

In addition to the methods listed, the study considered the possibility of using combined approaches, particularly integrating natural language processing methods with more complex self-learning models (e.g., BERT transformers). However, classical NLP approaches combined with machine learning were chosen to ensure the results' interpretability and optimal performance.

3.2. Sample

Three main blocks of data were collected for the study:

1. **Text data** – 500 threatening messages obtained from open sources: social networks Facebook, Instagram, Reddit public forums and Twitter microblogs. Keywords and phrases that potentially indicate aggressive content were used to select texts, such as "threat", "blackmail", "you will regret it", etc.

Semantic analysis and automatic extraction of texts containing keywords were used to select messages. However, it is worth noting that a specific part of aggressive content may remain unnoticed due to the peculiarities of filtering algorithms and user language structure changes to bypass automatic control systems.

2. **Behavioural data** – information about the online activity of 100 anonymized users who demonstrate risky behaviour patterns. This data was collected with the users' consent as part of public monitoring initiatives.

Behavioural pattern data was obtained with user consent and anonymized before analysis. A sample of 100 profiles was formed to ensure representativeness and coverage of different types of social activity, allowing for assessing risk behaviour patterns in a broad context.

3. **Human rights defenders** – a sample of 20 people involved in an online survey. The respondents were selected from about 100 human rights defenders through professional networks, including LinkedIn and specialized associations. A targeted approach was used for selection, taking

into account experience in the field of digital law enforcement for over three years and participation in the investigation of crimes in the digital environment.

The number of respondents, 20, was determined based on an expert selection criterion, which ensures a qualitative analysis of their assessments. The study involved human rights defenders with at least three years of experience working with digital threats, increasing the results' validity.

The amount of text and behavioural data was selected taking into account the need to ensure statistical reliability of the results. The sample of human rights defenders was formed to obtain expert assessments of the AI use.

It is essential to consider that sample size may affect the generalizability of results, so further research may focus on expanding the data volume and involving a more significant number of respondents.

3.3. Tools

The following tools were used for the study:

- Google Colab – to perform text data analysis using NLP libraries (spaCy, NLTK).
- Scikit-learn – to simulate behavioural patterns.
- Google Forms – to conduct a survey of human rights defenders.
- SPSS – to process and analyse survey results.

SpaCy and NLTK libraries were chosen for text data analysis because they provide flexibility in semantic text analysis and allow efficient work with large data sets. Machine learning algorithms were implemented using Scikit-learn, which allowed for comparing different approaches to behavioural risk modelling.

The use of these tools ensured the integration of qualitative and quantitative approaches, which allowed obtaining accurate and reliable results.

Using a comprehensive approach to data analysis allowed us to obtain results with a high level of reliability. However, it is necessary to consider that possible errors in determining risks may be associated with the peculiarities of natural language processing and the limitations of the selected algorithms.

4. RESULTS

To achieve the goal of the study, the collected text data was analysed, which has a heterogeneous structure: from short, concise messages on Twitter to detailed and emotionally rich texts on Reddit. The analysis of this sample makes it possible to study how aggressive content manifests itself in different formats and environments. This provides gain deeper insights into the types of threats and ways to identify them.

Textual data analysis reveals that the nature of threatening content depends on the platform. Short ultimatums or sarcastic threats are most common on Twitter due to message length limitations. At the same time, Reddit allows users to post longer texts, which facilitates a detailed presentation of threatening intentions or manipulative strategies of psychological pressure. Similar differences between platforms have been noted in previous studies, confirming the importance of analyzing the social context for detecting digital threats.

Table 1 shows the number of messages collected from each platform, as well as their percentage contribution to the total sample.

Table 1: The Number Of Messages Collected By Platform

Platform	Number of messages	Percentage (%)
Facebook	200	40
Instagram	100	20
Reddit	120	24
Twitter	80	16

Source: developed by the authors based on the results of their own research.

The text data collected in the study were classified into four main types of threatening messages: explicit threats, implicit threats, blackmail, and cyberbullying. The classification was based on semantic analysis of the content of the texts using predefined criteria.

Their frequency explains the choice of these four categories in previous studies of threatening content and their importance for automated analysis. Explicit threats are often direct statements that intend to cause harm, making them easy for NLP algorithms to detect. Latent (hidden) threats, on the other hand, have an indirect form

and depend on the context, which creates difficulties for automated analysis. For example, phrases like "Remember what happened to others" or "I would think twice if I were you" may not contain an explicit threat but, in a particular context, are interpreted as manipulative pressure.

Explicit threats include messages that contain an immediate threat of physical or psychological harm, such as: "I will find you" or "You will regret doing this." Latent threats are less obvious, but can be perceived as a hint of danger, for example: "Think twice before doing this." Blackmail involves pressure or demand, backed up by a threat to disclose information or damage reputation: "If you don't do this, I will tell everyone." Cyberbullying encompasses systematic harassment and humiliation directed at the victim using text messages.

The classification of threatening messages is not always unambiguous, as some messages could combine several types of threats, for example, latent threats combined with elements of blackmail. This required additional verification of the accuracy of the distribution by category and analysis of possible classification errors.

Figure 1 presents the distribution of types of threatening messages in the sample.

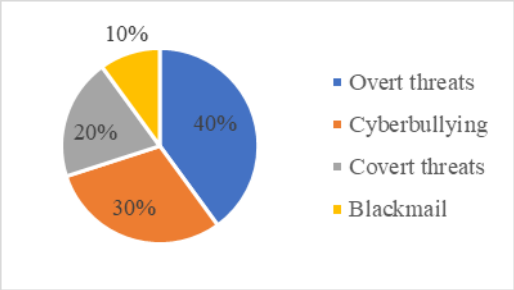

Figure 1: Distribution Of Types Of Threatening Messages In The Sample
Source: Developed By The Authors Based On The Results Of Their Own Research

The analysis of 500 messages revealed that overt threats made up the largest proportion of the sample (40%), while cyberbullying was the second most common type of threat (30%). Covert threats made up 20%, and blackmail was the smallest (10%). These results suggest that overt and systematic forms of aggression are the most common in the selected environment.

Explicit threats dominate other messages because they are the easiest to express in text form and do not require additional conditions, such as blackmail, which usually requires compromising information. Cyberbullying also occupies a significant share of the sample because the online environment often promotes anonymity, which, in turn, lowers social barriers to aggressive behaviour.

The analysis of text data also allowed us to identify the most frequently used words and phrases that are often found in threatening messages. These keywords are markers of aggression and threats used for psychological pressure or intimidation. The most common expressions are "threat", "you'll regret it", "I'll tell everything", "Do it differently."

Analysis of platforms revealed differences like threatening content. Facebook and Reddit showed the highest frequency of aggressive language use, which may be because these platforms are widely used for discussions and debates, where conflicts often arise. With its limited message length, Twitter is more conducive to emotional, concise threats, while Instagram is focused on visual content, and media files often accompany threats.

The frequency of their use varies depending on the platform. For example, the words "threat" and "you'll regret it" were most often found in messages from Facebook and Reddit, which indicates the more aggressive nature of discussions on these platforms. The expression "I'll tell everything" was more popular on Twitter, where messages are usually short and emotionally charged, while "Do it differently" was more common on Instagram, where this expression accompanied visual content.

It is worth noting that the frequency of certain words can vary depending on their specific context. While some words are clear markers of threat, their exact meaning and level of aggression depend on how they were written and the tone of the discussion.

Table 2 shows the top 10 keywords and phrases that were most often found in messages, indicating their frequency and the platform on which they were most often used.

Table 2: Top 10 Keywords And Phrases In Threatening Messages

Key word/phrase	Frequency of use	The highest frequency platform
threat	50	Facebook
You will regret it	45	Reddit
I will tell everything	30	Twitter
Do it differently	25	Instagram
think twice	20	Facebook

I know your place	15	Reddit
you will not hide anymore	15	Instagram
it will end badly	10	Twitter
we will find you	10	Facebook
you are in trouble	5	Reddit

Source: Developed By The Authors Based On The Results Of Their Own Research

The analysis of the effectiveness of natural language processing (NLP) algorithms in the study was carried out by comparing the results of automated text analysis with the results of manual analysis performed by moderators. The NLP algorithms automatically processed 500 collected messages, classifying them by threat level. The moderators, in turn, performed a manual assessment using the same classification criteria.

Analysis of the results of NLP algorithms showed that although their accuracy is high (85%), they still make a significant number of errors, particularly in cases where the threat is veiled or contains specific language constructs. Although manual analysis demonstrates higher accuracy (95%), it requires significant human resources, which limits its application in large-scale studies.

The results showed that the accuracy of the NLP algorithms was 85%, while the manual analysis provided an accuracy of 95%. The errors of the algorithms were divided into two main categories: false positives (when a neutral message is mistakenly classified as threatening) and false negatives (when a threatening message was not identified). False positives in the NLP work were 10%, while false negatives were 5%.

The main reason for errors is that algorithms can misinterpret context, especially in sarcastic statements or implicit threats. For example, the statement "Yes, of course, I'll 'find' you..." may not be perceived as a threat by the algorithm due to the lack of explicit threatening markers, although such context would be evident to a human.

Table 3 presents a comparison of the accuracy of text analysis performed by the NLP algorithms and moderators.

Table 3: Comparison Of The Accuracy Of NLP Analysis And Manual Moderator

Method	Accuracy (%)	False positives (%)	False negatives (%)
NLP analysis	85	10	5
Manual analysis	95	3	2

Source: Developed By The Authors Based On The Results Of Their Own Research

The results indicate that NLP algorithms have significant potential for automating the analysis of threatening content, but their accuracy needs to be improved to minimize misclassifications. Combining such algorithms with manual analysis can provide a more effective approach to detecting threats in large volumes of text data.

Despite the significant potential of NLP algorithms in threat detection, their effectiveness depends on the accuracy of identifying threatening messages. Using such technologies in law enforcement requires additional testing and adaptation to real-world scenarios to avoid misclassification risks.

ML algorithms such as Decision Trees, Random Forest, and SVM were used to analyse behavioural patterns. They identified dependencies between users' behavioural characteristics and potential risks of aggressive or criminal behaviour.

Machine learning algorithms have revealed correlations between users' behavioural characteristics and possible aggressive or criminal behaviour risks. In particular, the analysis showed that certain communication features, such as the frequency of use of aggressive language or active participation in conflict discussions, can indicate increased risk.

The data from 100 anonymized profiles from Facebook, Instagram, Reddit, and Twitter provided a wide range of information on publication frequency, use of aggressive language, and participation in conflict discussions, which was used to train the algorithms.

The sample of 100 anonymized profiles was formed to cover users from different social networks, giving us a broader picture of behavioural patterns. However, this sample may not reflect the full range of possible behavioural scenarios since individual communication characteristics can vary significantly depending on the context and social environment.

Table 4 lists the sources and key characteristics of the data that were used to model behavioural patterns.

Table 4: Data Sources And Structure For Modelling

Data source	Number of profiles	Key indicators
Facebook	30	Likes, comments, post frequency
Instagram	20	Likes, photo captions
Reddit	30	Discussion topics, replies
Twitter	20	Tweets, replies

Source: Developed By The Authors Based On The Results Of Their Own Research

The data were collected from various sources in order to identify behavioural patterns and assess the accuracy of the modelling. The algorithms took into account both simple and complex dependencies between variables.

Algorithms were used for modelling that allows for direct correlations between behavioural traits and more complex dependencies. However, the accuracy of the analysis depends on the context of the messages since the exact words can have different meanings depending on the situation.

Key trends indicating a tendency to aggressive or potentially criminal activity include frequent use of aggressive language, participation in conflict discussions, and publication of provocative content. These patterns have become the main indicators of risk.

The main signs of risky behaviour were the frequent use of aggressive language, involvement in conflict discussions, and publication of content that could provoke hostility. However, such indicators do not always indicate real threats, as some users may use a conflict communication style without intending to commit offences.

Additional indicators include regular arguments, a hostile tone of messages, and rhetorical questions with a threatening subtext, such as: “Do you think I'm just going to leave this like that?” or “What will you do when everyone finds out?”. These markers effectively detect aggressive communication.

Additional characteristics that may indicate risky behaviour include regular arguments, a hostile tone of messages, and rhetorical questions with a hidden hint of threat. For example, phrases like “Do you think I’m just going to leave it like this?” or “What will you do when everyone finds out?” can be used as markers of aggressive communication. However, their interpretation depends on the overall context of the dialogue, which creates specific difficulties for automated analysis.

Figure 2 presents the distribution of the main risk indicators in user profiles, which allows us to visualize the frequency of their appearance.

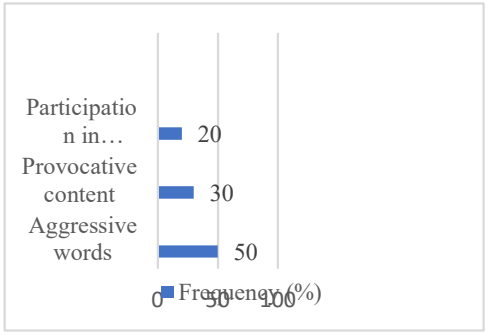

Figure 2: Distribution Of Key Risk Indicators In User Profiles
Source: Developed By The Authors Based On The Results Of Their Own Research

The analysis of these indicators showed their frequency among the 100 analysed profiles. The most common indicator was aggressive words, which were found in 50% of profiles. Provocative content ranked second (30%), while participation in conflicts was 20%.

The accuracy of the modelling was assessed by comparing the predictions of ML algorithms (Decision Trees, Random Forest, SVM) with the actual user actions determined by moderators. The indicators of accuracy (Precision), completeness (Recall) and F1-Score helped to assess the ability of the algorithms to correctly identify risks. Decision Trees provided an accuracy of 75%, Random Forest – 85%, and SVM – 90%, demonstrating the best results in complex cases.

Table 5 presents the results of evaluating the effectiveness of the three algorithms by the main metrics.

Table 5. Evaluation Of The Effectiveness Of Behaviour Modelling Algorithms

Algorithm	Accuracy (%)	Recall (%)	F1-Score (%)
Decision Trees	75	70	72
Random Forest	85	80	82
Support Vector Machine (SVM)	90	85	87

Source: Developed By The Authors Based On The Results Of Their Own Research

The results show that Random Forest and SVM models provide high efficiency in risk detection, but require significant computing resources. Decision Trees are a simpler and faster option, but are inferior in terms of accuracy and reliability.

Analysis of the results showed that the Random Forest and SVM algorithms demonstrate the highest efficiency in risk detection, but their use requires significant computational resources. In

contrast, the Decision Trees model is faster in execution but inferior in accuracy, especially in complex dependencies between variables. Additional analysis showed that the accuracy of the algorithms can vary depending on the input data type. For example, SVM turned out to be the most effective in recognizing latent threats, while Decision Trees cope better with classifying explicit threats.

The results of the survey of human rights defenders showed a variety of views on the AI use in detecting crimes against the will, honour, and dignity of a person.

The survey of human rights defenders showed a wide range of opinions on using AI to detect crimes against the person's will, honour and dignity. Although 50% of respondents fully support the use of AI, 30% express partial support, pointing to the risks of misuse of the technology. Among the main concerns, respondents noted the possibility of discrimination and bias in algorithms (45%), the threat to privacy (30%), and the lack of proper state regulation (25%). At the same time, 40% of those surveyed believe that AI can significantly speed up the process of analyzing threatening content and increase the efficiency of law enforcement agencies.

Figure 3 presents the results of the survey of human rights defenders regarding their attitude to the AI use in detecting crimes.

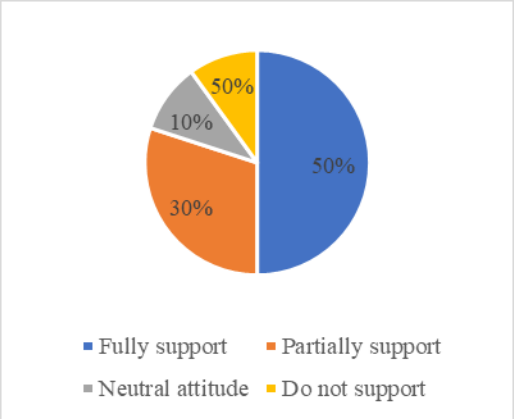

Figure 3: Attitudes Of Human Rights Defenders Towards The AI Use In Crime Detection
Source: Developed By The Authors Based On The Results Of Their Own Research

Human rights defenders' attitudes towards the AI use in crime detection are mostly positive. Half of the respondents (50%) fully support the AI use, recognizing its ability to provide speed and efficiency of analysis. Another 30% partially support the implementation of these technologies, although they express some reservations about possible risks. A neutral attitude is observed among 10% of participants, while another 10% oppose the use of AI because of the risks of human rights violations and potential abuses.

The survey results showed a variety of opinions on the use of AI in detecting crimes against the will, honour, and dignity of a person. 60% of respondents have experience in human rights protection for more than three years, and 70% have encountered violations of rights in the digital environment, such as cyberbullying or blackmail.

The advantages of AI include the speed of analysis (40%), accuracy in detecting threats (30%), and automation (20%). The main disadvantages were the possibility of errors (35%), risks to privacy (30%) and insufficient transparency of algorithms (20%).

The assessments of the ethics of using AI were mixed: 60% supported it if it met the rules, 20% linked acceptability to context, and 20% did not support its implementation. In general, 70% of respondents indicated the need for strict regulation, while 20% advocated for general ethical principles.

Human rights defenders also proposed a number of measures to improve the effectiveness of AI. The need for training specialists (40%), improving algorithms (35%), and developing new ethical standards (25%) were the most frequently mentioned. This emphasizes the importance of combining technical improvements in technology with ensuring respect for human rights.

In summary, human rights defenders' attitudes towards AI align with global trends. Similar studies indicate a high level of interest in implementing AI to combat online threats but also confirm the need for enhanced regulation. Thus, the results highlight the need to develop balanced approaches to using AI in the context of human rights protection.

5. DISCUSSION

The analysis found both confirmation and discrepancies with earlier studies, which reveals the potential and limitations of using AI in combating crimes against the will, honour, and dignity of a person. The evaluation results of artificial intelligence technologies' effectiveness confirm their high accuracy in detecting threatening content and predicting criminal behaviour. At the same time, the practical application of these technologies faces challenges, including ethical issues,

regulatory issues, and the accuracy of automated classification.

According to [15], it is crucial to consider social and psychological contexts when implementing AI, especially in the field of mental health. This approach is consistent with findings suggesting that considering users' behavioural and cultural characteristics contributes to more accurately identifying potential risks. At the same time, the findings indicate the need for additional verification mechanisms to minimize misclassifications.

Wang and Ma [16] note that ML algorithms are effective in preventing public health crimes, emphasizing the importance of integrating technology into overall risk management strategies. Analysis of the algorithms' accuracy in predicting risks confirms these findings, but further improvement of AI models is necessary to adapt to different crime scenarios and reduce the likelihood of incorrect classification.

Research by Saini and Kaur [17] highlights the efficiency of predictive analytics in identifying high-risk areas. While their work primarily focuses on spatial analysis, other sources indicate that the reliability of results depends not only on geographical data but also on textual and behavioural indicators. The results show that analysing behavioural patterns and user activity on social networks is essential for effective threat detection, confirming the significance of behavioural factors in risk prediction.

The authors in [18] introduced the concept of an intelligent policing system based on large language models. This supports the conclusions regarding the effectiveness of NLP algorithms in identifying threatening messages. However, the adaptability of algorithms to context remains a key challenge. The results show that NLP models demonstrate high accuracy in classifying threatening messages. Still, the effectiveness of these algorithms largely depends on the linguistic features and communication practices of users of different platforms.

Yen and Hung [19] emphasize the necessity of social justice and transparency in AI-driven crime prediction. Comparing the obtained assessments with these conclusions confirms that the risks of algorithmic bias and the possibility of discrimination are serious challenges that require attention. A significant number of the respondents emphasize the need to improve legal regulation and ensure transparency of the work of algorithms, which emphasizes the relevance of developing ethical standards for using AI in law enforcement.

Fors [20] draws attention to the risks of data manipulation and lack of transparency in algorithms, which is also reflected in many other studies. This emphasizes the importance of ethical regulation and adherence to standards in the development of AI systems. The results confirm the need to develop clear ethical standards for the use of AI, especially in the context of processing threatening content. The lack of transparency in algorithm operation can decrease trust both among users and within law enforcement practice.

Catalina et al. [21] focus on public awareness of the use of AI in medical research. Their findings confirm that a lack of understanding of the principles of algorithmic work among the public can lead to mistrust, which is in line with common challenges in other fields, including human rights. These findings are consistent with the challenges of applying AI to digital security, where low awareness about how algorithms work can become a barrier to their implementation and legitimacy.

The researchers in [22] examine the barriers and opportunities for the AI implementation in the medical field, emphasizing the importance of ensuring transparency and accountability of algorithms. These findings correlate with the general need for ethical regulation of the AI use, which has been noted by many authors. Similar challenges are observed in cyber threat detection, as insufficient accountability and the difficulty of verifying algorithm decisions can create risks of wrongful accusations or systematic errors in classification.

Sapignoli [23] focus on the threats associated with global data governance, which often precedes regulatory action. This complements existing discussions on the need to develop international standards for the AI use in law enforcement. The lack of uniform international standards for the use of AI in the security sector creates additional risks, as confirmed by both empirical research and surveys of human rights activists, which indicate the need for strict regulation of algorithmic decisions in law enforcement.

Woo et al. [24] compare approaches to regulating digital health technologies in the US and Korea, emphasizing the need to align local and international standards. Similar challenges arise in law enforcement technologies, where adapting algorithms to different legal systems is key. The data confirms that adapting AI to law enforcement mechanisms is complex, as it requires considering

legal, social, and ethical norms in each specific jurisdiction.

Kavanagh et al. [25] examine the risks of violence in healthcare, analysing the potential benefits and drawbacks of using AI. The findings on the importance of rigorous testing of algorithms to avoid erroneous decisions can be extrapolated to the field of predicting criminal behaviour. Similarly, in combating crimes against the will, honour, and dignity of the individual, the need for careful verification of algorithms has been identified to avoid excessive reliance on automated decisions, especially in cases of high social significance.

Bayerl et al. [26] conduct a cross-country comparison of AI strategies in law enforcement agencies in Greece, Italy, and Spain. Their results emphasize the importance of cultural context for the successful implementation of algorithms, which is also reflected in human rights practices. Cultural context analysis has also proven critical when studying the implementation of AI in law enforcement, as algorithms must consider the specifics of the communication environment to ensure their correct operation.

Evans [27] analyses the legal aspects of AI use, emphasizing that medical algorithms often do not take into account complex social factors. This reminds us of the risks of a simplistic approach to implementing technologies in law enforcement, where consideration of individual circumstances is critical. The author points out that the problem of algorithm universality is also relevant in law enforcement since general automation without adaptation to specific conditions can reduce the effectiveness of technologies and cause additional legal risks.

Malhotra and Misra [28] examine the issue of accountability of algorithms in decision-making, emphasizing the need for a clear regulatory framework. These findings are relevant to the field of crime prediction, where a lack of transparency in algorithms can lead to bias or discrimination. The results confirm the importance of developing regulations to regulate AI algorithms, as the lack of clear rules can contribute to bias and the risk of human rights violations during automated decision-making.

The researchers in [29] examine the issue of value alignment when designing forensic algorithms, emphasizing the importance of taking into account different cultural approaches. This is consistent with the need to adapt algorithms to different social contexts. It has been confirmed that AI algorithms' effectiveness largely depends on the social and cultural context. Algorithms that do not consider the peculiarities of linguistic communication and national traditions may demonstrate reduced accuracy and increased false positives.

Srikanth and Sowmya [30] and Zhu and Zheng [30] analyse the impact of AI on the judicial system, emphasizing the need for clear rules for its implementation. Their conclusions emphasize that the use of such technologies requires a careful and balanced approach. Human rights activists also emphasize the need to develop specific protocols for using AI that are consistent with the conclusions about the need for a balance between automation and human control in judicial and law enforcement practice.

In general, a comparison with other works shows that the AI use has significant potential in various areas, but requires careful regulation, taking into account the social context and ensuring transparency of algorithms. Various studies' analyses confirm that while AI can significantly improve the effectiveness of threat detection and prediction, its implementation must be accompanied by appropriate control mechanisms and adaptation to specific social realities.

5.1. Limitations

The study is limited by the availability of text data from social networks, which may affect the accuracy of modelling behavioural patterns. The limited number of respondents (20 human rights defenders) reduces the representativeness of the obtained results. Besides, only basic NLP algorithms were used without the use of more complex self-learning models, which could affect the accuracy of the analysis.

Another limitation of the study is that the analysis of threatening messages was based on open sources, which means that some potentially important information may have remained unavailable due to users' privacy settings or platform policies. In addition, the sample of human rights defenders, although formed from specialists working with digital threats, may not fully reflect the full range of opinions of the expert community. Using basic NLP algorithms without deep learning also creates the risk of insufficient adaptation of the models to more complex linguistic constructs and contextual features of threats.

5.2. Recommendations

For the effective implementation of AI in combating crimes against freedom, honour, and dignity, it is important to develop standardized

protocols that take into account ethical and legal aspects, and to establish international cooperation for data exchange between law enforcement agencies. It is recommended to create educational programmes for human rights defenders and law enforcement officers, as well as initiatives to raise public awareness of the possibilities and limitations of AI.

Particular attention should be paid to creating comprehensive mechanisms for verifying decisions made by AI algorithms to prevent possible misclassifications or discrimination against certain groups of users. To increase the effectiveness of technologies, multi-level machine learning models should be expanded, allowing for better recognition of complex language patterns and the context of threats. It is also essential to regularly monitor the accuracy of AI and adapt the models to new challenges in the digital environment. It is recommended that the interdisciplinary approach that combines technical analysis with human rights and social aspects be strengthened to develop more effective methods for protecting users.

6. CONCLUSIONS

The results of the study demonstrate the significant potential of AI in combating crimes against the will, honour, and dignity of a person. NLP algorithms have shown high efficiency in identifying threatening messages, ensuring the accuracy of analysis even with large amounts of data. Modelling of behavioural patterns allowed to identify key risks, in particular the use of aggressive content and participation in conflict discussions, which can become the basis for predicting criminal behaviour. A sociological survey of human rights defenders confirmed the importance of ethical regulation and transparency during the implementation of AI, and also highlighted potential benefits, such as automation of routine processes and reducing the burden on the human factor.

However, the results highlight the need to improve algorithms, especially in reducing false positives and considering the context of messages. Increasing the effectiveness of AI is possible through using complex deep learning models and adapting algorithms to users' linguistic and cultural characteristics. The issue of transparency of decision-making mechanisms by algorithms remains essential, and it requires the development of appropriate standards for evaluating and verifying the results of automated threat analysis.

The academic novelty of the study is the integrated approach to analysing the effectiveness of AI, which combines text analysis, prediction of behavioural risks, and a sociological aspect. A feature of the study is the use of various social media platforms, which provide a broad context for analysis.

The study's scientific novelty lies in its interdisciplinary approach, which combines natural language processing methods, behavioural modelling, and sociological analysis. This allows for a more comprehensive risk assessment and the integration of technological solutions into the broader context of human rights activities.

The practical value of the research is the possibility of implementing the obtained results in law enforcement activities, in particular to increase the accuracy of threat detection and build early warning systems. The obtained data can be used to develop recommendations on the ethical use of AI, create regulations and educational programmes for human rights defenders and law enforcement officers.

The practical value of the study lies in the possibility of using its findings to improve automated threat monitoring systems, as well as in the development of training programs for human rights defenders and law enforcement officers. The proposed approach can be adapted to different jurisdictions and applied in international practice to combat digital crimes.

REFERENCES:

[1] R. Shchokin, V. Oliinyk, O. Amelin, V. Maziychuk, and D. Kyslenko, “Methods of combating offenses in the field of ecology”, *Journal of Environmental Management and Tourism*, Vol. 14, No. 1, 2023, pp. 5-15. doi: 10.14505/jemt.v14.1(65).01.

[2] A. Rimo, “Is prevention better than cure? The ever-increasing criminalisation of acts preparatory to an offence in Spain”, *International Journal for Crime, Justice and Social Democracy*, Vol. 10, No. 1, 2020, pp. 1-14. doi: 10.5204/ijcjsd.v10i1.1502.

[3] A. Jin, and N. Salehi, “(Beyond) reasonable doubt: Challenges that public defenders face in scrutinizing AI in court”, *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery (United States), May 11, 2024, art. 467. doi: 10.1145/3613904.3641902.

[4] C. Wu, H. Xu, D. Bai, X. Chen, J. Gao, and X. Jiang, "Public perceptions on the application of artificial intelligence in healthcare: A qualitative meta-synthesis", *BMJ Open*, Vol. 13, No. 1, 2023, art. e066322. doi: 10.1136/bmjopen-2022-066322.

[5] K. Kieslich, and M. Lünich M., "Regulating AI-based remote biometric identification. investigating the public demand for bans, audits, and public database registrations", *Regulating AI-Based Remote Biometric Identification. Investigating the Public Demand for Bans, Audits, and Public Database Registrations*, Association for Computing Machinery (United States), June 3-6, 2024, pp. 173-175. doi: 10.1145/3630106.3658548.

[6] R. Broadhurst, D. Maxim, P. Brown, H. Trivedi, and J. Wang, "Artificial intelligence and crime", Korean Institute of Criminology, Australian National University, 2019. doi: 10.2139/ssrn.3407779.

[7] R. Hardy, K. Glastonbury, S. Onie, N. Josifovski, A. Theobald, and M. Larsen, "Attitudes among the Australian public toward AI and CCTV in suicide prevention research: A mixed methods study", *American Psychologist,* Vol. 79, No. 1, 2024, pp. 65-78. doi: 10.1037/amp0001215.

[8] S. Rankin, "Technological tethereds: Potential impact of untrustworthy artificial intelligence in criminal justice risk assessment instruments", *Washington and Lee Law Review*, Vol. 78, No. 2 (Spring 2021), 2020, pp. 647-724. doi: 10.2139/ssrn.3662761.

[9] X. Wang, Y. Wu, X. Ji, and H. Fu, "Algorithmic discrimination: Examining its types and regulatory measures with emphasis on US legal practices", *Frontiers Artificial Intelligence*, 7, 2024, 1320277. doi: 10.3389/frai.2024.1320277.

[10] I. Kopotun, A. Nikitin, N. Dombrovan, V. Tulinov, and D. Kyslenko, "Expanding the potential of the preventive and law enforcement function of the security police in combating cybercrime in Ukraine and the EU", *TEM Journal*, Vol. 9, No. 2, 2020, pp. 460-468. doi: 10.18421/TEM92-06.

[11] R. Adam, D. Schwartz, S. Sanga, A. Charlotte, K. Hammond, L. Amaral, and S. Consortium, "The promise of AI in an open justice system", AI Magazine, Vol. 43, No. 1, 2022, pp. 69-74. doi: 10.1609/aimag.v43i1.19127.

[12] E. Terhune, P. Heyn, C. Piper, C. Wethey, A. Monley, M. Cuevas, and N. Miller, "Association between genetic polymorphisms and risk of adolescent idiopathic scoliosis in case-control studies: a systematic review", *Journal of Medical Genetics*, 2023. doi: 10.1136/jmg-2022-108993.

[13] T. Hubanova, R. Shchokin, O. Hubanov, V. Antonov, P. Slobodianiuk, and S. Podolyaka, "Information technologies in improving crime prevention mechanisms in the border regions of southern Ukraine", *Journal of Information Technology Management*, Vol. 13, 2021, pp. 75-90. https://doi.org/10.22059/jitm.2021.80738.

[14] T. Kortukova, Y. Kolosovskyi, O. Korolchuk, R. Shchokin, and A. Volkov, "Peculiarities of the legal regulation of temporary protection in the European union in the context of the aggressive war of the Russian Federation against Ukraine", *International Journal for the Semiotics of Law*, Vol. 36, 2023, pp. 667-678. doi: 10.1007/s11196-022-09945-y.

[15] S. Cross, I. Bell, J. Nicholas, L. Valentine, S. Mangelsdorf, S. Baker, N. Titov, and M. Alvarez-Jimenez, "Use of AI in mental health care: Community and mental health professionals survey", *JMIR Mental Health*, Vol. 11, 2024, art. e60589. doi: 10.2196/60589.

[16] H. Wang, and S. Ma, "Preventing crimes against public health with artificial intelligence and machine learning capabilities", *Socio-Economic Planning Sciences*, Vol. 20, 2021, art. 101043. doi: 10.1016/J.SEPS.2021.101043.

[17] I. Saini, and N. Kaur, "The power of predictive analytics: forecasting crime trends in high-risk areas for crime prevention using machine learning", *14th International Conference on Computing Communication and Networking Technologies (ICCCNT),* IEEE (India), July 06-08, 2023. doi: 10.1109/ICCCNT56998.2023.10306731.

[18] P. Sarzaeim, Q. Mahmoud, and A. Azim, "A framework for LLM-assisted smart policing system", *IEEE Access*, Vol. 12, 2024, pp. 74915-74929. doi: 10.1109/ACCESS.2024.3404862.

[19] C. Yen, and T. Hung, "Achieving equity with predictive policing algorithms: A social safety net perspective", *Science and Engineering Ethics,* Vol. 27, 2021, art. 36. doi: 10.1007/s11948-021-00312-x.

[20] K. Fors, "Predict and surveil: Data, discretion, and the future of policing, by Sarah Brayne: A review by Karolina La Fors", *Information*

Polity, Vol. 26, No. 3, 2021, pp. 327-330. doi: 10.3233/ip-219008.

[21] Q. Catalina, J. Femenia, A. Fuster-Casanovas, F. Marin-Gomez, A. Escalé-Besa, J. Solé-Casals, and J. Vidal-Alaball, "Knowledge and perception of the use of AI and its implementation in the field of radiology: cross-sectional study", Journal of Medical Internet Research, Vol. 25, 2023, e50728. doi: 10.2196/50728.

[22] E. Lee, J. Torous, M. Choudhury, C. Depp, S. Graham, H. Kim, M. Paulus, J. Krystal, and D. Jeste, "Artificial intelligence for mental healthcare: clinical applications, barriers, facilitators, and artificial wisdom", *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, Vol. 6, No. 9, 2021, pp. 856-864. doi: 10.1016/j.bpsc.2021.02.001.

[23] M. Sapignoli, "The mismeasure of the human: Big data and the 'AI turn' in global governance", *Anthropology Today*, Vol. 37, No. 1, 2021, pp. 4-8. doi: 10.1111/1467-8322.12627

[24] J. Woo, E. Kim, and S. Kim, "The current status of breakthrough devices designation in the United States and innovative medical devices designation in Korea for digital health software", *Expert Review of Medical Devices*, Vol. 19, No. 3, 2022, 213-228. doi: 10.1080/17434440.2022.2051479.

[25] K. Kavanagh, C. Pontus, and L. Cormier, "Healthcare violence and the potential promises and harms of artificial intelligence", *Journal of Patient Safety*, Vol. 20, No. 5, pp. 307-313. doi: 10.1097/PTS.0000000000001245.

[26] P. Bayerl, B. Akhgar, E. Mattina, B. Pirillo, I. Cotoi, D. Ariu, M. Mauri, J. Garcia, D. Kavallieros, A. Kardara, K. Karagiorgou, "Strategies to counter artificial intelligence in law enforcement: Cross-country comparison of citizens in Greece, Italy and Spain", arXiv:2405.19970, 2024, pp. 1-3. doi: 10.48550/arXiv.2405.19970.

[27] B. Evans, "Rules for robots, and why medical AI breaks them", *Journal of Law and the Biosciences,* Vol. 10, No. 1, 2024, pp. 1-35. doi: 10.1093/jlb/lsad001.

[28] P. Malhotra, and A. Misra, "Accountability and responsibility of artificial intelligence decision-making models in Indian policy landscape", *CEUR Workshop Proceedings*, Vol. 3215, 2022, art. 15. https://doi.org/10.4018/979-8-3693-5415-5.ch007.

[29] C. Winter, N. Hollman, and D. Manheim, "Value alignment for advanced artificial judicial intelligence", *American Philosophical Quarterly,* Vol. 60, No. 2, 2023, pp. 187-203. doi: 10.5406/21521123.60.2.06.

[30] B. Srikanth, and V. Sowmya, "Assessing artificial intelligence in judicial system", *International Artificial Intelligence in Judicial System*, Vol. 08, No. 07, pp. 1-5. 2024. doi: 10.55041/ijsrem36601.

[31] K. Zhu, and L. Zheng, "Based on artificial intelligence in the judicial field operation status and countermeasure analysis", *Mathematical Problems in Engineering*, Vol. 2021, 2021, art. 9017181. doi: 10.1155/2021/9017181.