DOI: 10.54264/0079

PATRYK ORGANIŚCIAK
Teaching and Research Assistant, Department of Complex Systems, The Faculty of Electrical and Computer Engineering, Rzeszow University of Technology; *e-mail: org@prz.edu.pl*; ORCID: 0000-0002-5277-4038

PAWEŁ KURAŚ
Teaching and Research Assistant, Department of Complex Systems, The Faculty of Electrical and Computer Engineering, Rzeszow University of Technology; *e-mail: p.kuras@prz.edu.pl;* ORCID: 0000-0002-8658-0821

BARTOSZ KOWAL
Teaching and Research Assistant, Department of Complex Systems, The Faculty of Electrical and Computer Engineering, Rzeszow University of Technology; *e-mail: b.kowal@prz.edu.pl;* ORCID: 0000-0002-7909-6484

s. 123-141

THREATS AND CRISIS EVENTS DETECTION USING MACHINE LEARNING TECHNIQUES WITH SOCIAL MEDIA DATA

ABSTRACT

In the paper, the authors present the outcome of web scraping software allowing for the automated classification of threats and crisis events detection. In order to improve the safety and comfort of human life, an analysis was made to quickly detect threats using a modern information channel such as social media. For this purpose, social media services that are popular in the examined region were reviewed and the appropriate ones were selected using the criteria of accessibility and popularity. Approximately 300 unique posts from local groups of cities and other administrative centers were collected and analyzed. The decision of which entry was classified as a threat was defined using the ChatGPT tool and the human expert. Both variants were tested using machine learning (ML) methods. The paper tested whether the ChatGPT tool would be effective at detecting presumed events and compared this approach to the classic ML approach.

KEYWORDS

social media, threat detection, web scraping, chatgpt, machine learning

1. INTRODUCTION

At present, the use of social media for information purposes is widespread [1]. The process of information propagation is performed between many actors, and despite the complex business models of the well-known social networks, there is communication not only in the form of advertiser-receiver, but also between users, i.e., receiver-receiver. Recipients communicate in many ways: publicly, privately, one-way or two-way [2]. Topics of information exchange include the presentation of companies' services [3, 4], local and global events [5, 6, 7], offers to buy and sell [8, 9] exchanges of opinions, or current threats that may adversely affect safety and living comfort [10]. Such a wide pool of information, which is often publicly available, is an excellent opportunity to attempt rapid detection of possible threats. The rapid detection of threats makes it possible to speed up the response to them and thus minimize their impact [11].

The theoretical aspect of the issue under investigation in the article determines an attempt to detect local threats based on the messages of social media users. Such threats may be, for example, car accidents or fires. It is a natural phenomenon that people connect and unite into local groups on social media. This allows them to support each other, to warn of road obstructions or a series of home burglaries through posts that can potentially be read by anyone but are targeted at those in the area. This means that even though the residents of a region are informed of an incident, the city management, road management or services such as the police or fire department are not necessarily informed. Potentially, it is possible to automatize the process of detection and notify of any events that require it. While information itself will not actively pose a challenge in the world of common communication, the detection of these events itself is troublesome. Making such attempts requires appropriate sets of user messages, which are worth imbuing the decision model with. Chat GPT is one of such models, but it is built on very general data, which can have good as well as bad effects. Another way is to create your own machine learning model on the basis of data only taken from social media. These models can have the ability to self-learn based on increasing entries and expert decisions on whether a threat actually occurred or it's just false positive message.

In this research work, we set out to look at the role of social media during threats, as well as explore specific machine learning methods that could help detect such events.

The motivation for this approach was the popularity of social media among people, easy access to data and the evolution of information propagation. Social media sites are an integral part of the modern world, so it is necessary to look for positive opportunities to exploit them. The proposed approach reduces the amount of work needed by humans and is universal for many regions.

The motivation for this paper is an attempt to solve a practical problem related to the rapid discovery of local threats. Rapid detection can mitigate the consequences of an incident and save people's health or lives. The proposed method can support the work of services such as the police as well as administrative districts.

This paper is organized as follows: in Section 2, we provide an overview of social media and its impact on society, highlighting both its positive and negative aspects. Section 2.1 delves into the negative impact of social media on social security issues, while Section 2.2 presents a case study of the positive use of social media during the 2023 earthquake in Turkey and investigates the potential of social media in threat detection. In Section 3, we discuss early warning methods and the role of the Government Security Center (RCB) in Poland, Central Reporting Application (CAR) and the National Map of Safety Threats. Section 4 is detailing methodology of research which results are presented and discussed in Section 5, with a focus on the effectiveness of machine learning methods in data analysis. The paper concludes with Section 6, which discusses the implications of the findings and potential future research directions.

2. SOCIAL MEDIA

Social media services are platforms for communication among Internet users, are an important part of modern society and the evolution of communication media [12]. Beginning with signals and language, through print and telecommunications, to the era of computerization and the Internet, social media services represent the latest stage in the evolution of communicators. With widespread access to constant wireless connections, Bluetooth and WiFi technology, as well as smartphone functionality, communication is becoming extremely easy and accessible anywhere. What makes social media unique from other online platforms is its two-way nature and the fact that it is the users themselves who create and shape the content on these sites. Rost et al in their study in 2013 [13] argues for the importance of analyzing social media, specifically Foursquare check-in data, as a communicative rather than representational system, highlighting the influence of user behavior and motivations on large data analysis. Thanks to this media revolution, every individual now has the right to a free and independent voice. Notifications, which are one of the key elements of social media, play an important role in the popularity of these sites, influencing the number of visitors and creating a positive opinion of the site's performance. Social media is a complex field that includes social networking sites, blogs and wiki-type sites [14, 15]. Social media had a massive impact in the context of the COVID-19 coronavirus pandemic among others in the sphere of trade, but also in social life [43].

2.1. The negative impact of social media on social security issues

The structure of the modern Internet is based on an implicit agreement: users are given free access to websites, apps and social media, while companies collect and sell their data to advertisers. This arrangement, however, appears to be spiraling out of control, with services such as Facebook, TikTok, Snapchat and YouTube accused of inciting violence, inadequately protecting children and fueling mental health problems [16]. In her TEDx talk titled „Is Social Media Hurting Your Mental Health?", Bailey Parnell discusses the profound impact of social media on mental health, drawing from both personal experience and empirical research [17]. She identifies four major social media stressors, where Highlight Reel refers to users' tendency to post only the best moments of their lives, leading to constant comparison with others [18]. Social Currency is the value placed on likes, comments and shares, which can significantly affect an individual's self-esteem and identity [19]. F.O.M.O (Fear of Missing Out), on the other hand, is social anxiety resulting from the fear of missing out on a potential connection, event or opportunity [20] Finally, online harassment, experienced by 40% of adult Internet users, is a significant stressor, especially for women, LGBTQ people and people of different skin color [21].

A common argument is the statement: „I have nothing to hide, so why should I worry?" [22, 23]. However, privacy is not about hiding secrets, but about controlling how information about us is collected, compiled and used [23]. This information can be used to make judgments that affect people's lives, such as qualifying to rent an apartment, buy a house or car, or get a job [22]. Data brokers play an important role by collecting, buying and selling private data without our knowledge or consent [24]. They can collect detailed information, including previous addresses, email addresses, recent purchases and even estimated body mass index and sleep quality [25]. Companies use algorithms and artificial intelligence to analyze this data and make meaningful decisions about our lives. Privacy abuse disproportionately affects marginalized communities [24].

Negative social media activity took place during the COVID-19 pandemic by increasing panic, among other things. An example of such behavior occurred in India. People's behavior led to stock of masks and sanitizers from the market and creating fake claims about transmission of virus and its survival on different surfaces which caused a rise of panic [26]. Research shows that 86.73% of respondents have experienced panic and it finds social media use has a significant impact on the development of panic among people over COVID-19 outbreak [27].

2.2. Positive social media use cases during emergencies: a case study of the 2023 earthquake in Turkey

In the digital age, social media has become an integral part of everyday life, and its role in emergency situations such as natural disasters is increasingly appreciated. This subsection presents an analysis of positive social media use cases during the 2023 earthquake in Turkey [28]. During the earthquake in Turkey, social media quickly became a major source of communication regarding search and rescue information. Local residents actively used social media platforms to share information about collapsed buildings, reports of signs of life under the rubble, and the need for help and support [29]. Social media played a key role in mobilizing communities to assist in search and rescue efforts. Many people used social media to share information about where professional rescue teams, equipment or heavy debris removal equipment were needed [28]. Social media allowed for first-hand information from those directly affected by the disaster. For example, a volunteer working in the rubble of one of the collapsed buildings asked a journalist to record him and share what he had to say about the situation on the ground [29]. On rare occasions, concerned residents have managed to interrupt live broadcasts from the affected areas, challenging the official narratives provided by controlled government news channels [28]. It should be remembered that despite all these positive aspects, social media has unfortunately also been used to „prey" on disaster victims and spread false information [30].

2.3. Threat detection using social media data

The occurrence of risks and the analysis of related information are not uncommon research subjects. The topics are also being discussed in relation to social media. Cox et al in their study in 2018 [31] that addresses risk management in the supply chain noted that real-time information is critical and enables an organization to make more informed and timely decisions on how to manage or mitigate risks. As an example, it cited the occurrence of a disaster that happened near a manufacturing plant, where information about the event can change planned travel routes. The study describes how social media significantly helps distribute valuable information. In the past Alexander reviewed the actual and potential use of social media in emergencies, disasters and calamities [32]. His work explores the various uses of social media in emergency and crisis situations, including monitoring situations, extending emergency response, crowdsourcing, creating social cohesion, and enhancing research, while also acknowledging potential negative impacts such as spreading rumors and undermining authority. It emphasizes the need for emergency managers to adapt to the increasing use of social media, while also considering ethical implications to prevent misuse during crises. Also, the paper written by Chen, Vorvoreanu & Madhavan in 2014 [33] proposes a workflow integrating qualitative analysis and data mining techniques to analyze social media data and understand engineering students' learning experiences, with a focus on Twitter posts, revealing issues such as study load, lack of social engagement, and sleep deprivation.

Social media can be a good data set for detecting different types of threats. In 2021 López-Vizcaíno, M. F. et al [34] demonstrated that such data can help in early detection of cyberbullying on social media networks.

Given the high popularity of social media, almost any person can immediately inform others directly of a threat, even without being a government employee or someone affiliated with public media such as radio or television. In addition, information entered in this way can be quickly confirmed or disproven by other users through comments. In the case of state media or alert systems, there is no opportunity for debate about reliability, because the message is one-way. Another benefit of the approach is the local nature of information sharing, so the message about potential threats reaches only those interested in them.

3. EARLY WARNING METHODS AND THREAT REPORTING ON THE EXAMPLE OF POLAND

Early warning systems are the tools needed to effectively deal with crisis situations that may affect the public. There are many approaches to early warning in the context of public opinion on the Internet that are used in research. Wang et al in ther work [35] used a fuzzy comprehensive evaluation method to predict public opinion online. Their early warning system considered the weight of each indicator and predicted the warning level. Cao et al in 2019 [36] constructed an early warning level model for public opinion using the analytic hierarchy process (AHP) and the fuzzy comprehensive evaluation method. The AHP process is one of the so-called multi-criteria decision-making methods and helps reduce inconsistency in the weights of the criteria selected, as Mazurek et al in 2021 and Koczkodaj et al in 1993 wrote about [37, 38]. Some early warning methods can be imperfect, especially when it comes to predicting phenomena with high uncertainty and randomness, such as public opinion. In such situations, artificial neural networks can prove more effective. Zhao & Pan in 2020 used BP (Back-Propagation) neural networks to construct an online public opinion early warning model [39]. Their results showed that the accuracy of public warning has improved, although there is still room for improvement.

3.1. Government Security Center (RCB) and RCB Daily Reports

The Government Security Center (pl. Rządowe Centrum Bezpieczeństwa - RCB) is the key body in Poland responsible for coordinating activities related to crisis management and national security. As part of its activities, the RCB generates various types of reports and alerts to inform relevant authorities and the public about potential threats and emergencies [40].

One of the RCB's key information products is the daily reports. These are published daily and contain the most important security information in the country for the past 24 hours. These reports can include information on natural threats, such as floods or fires, but also on public security threats, such as terrorist attacks or outbreaks [40]. Another important output is the RCB's analytical bulletins. These are created based on the analysis of various information sources and are designed to provide detailed analysis on specific threats or emergencies [40]. The RCB also works with telecommunications network operators to notify end users of threats. This is particularly important in situations where rapid and effective notification of the public can help minimize the impact of a potential crisis [41].

3.2. Central Reporting Application (CAR)

The Central Reporting Application (pl. Centralna Aplikacja Raportująca - CAR) is a tool created to improve crisis communication and reporting. CAR is a standard designed to improve communication related to emergency reporting and applies to a large number of units. The Central Reporting Application is the first automated emergency information flow tool that can cover everyone involved in emergency management [42]. The application was launched at provincial emergency management centers on September 1, 2013. It is currently being used at provincial and district emergency management centers and at the central level. CAR remains in use through 2023 and continues to expand with new functionalities. These include integration with existing provincial applications and systems, as well as inclusion of other central offices key to emergency management in reporting. New functionalities also include a module for automating the collection of data on victims, casualties, and losses to infrastructure and property.

3.3. National Safety Threats Map

The National Security Threat Map (pl. Krajowa Mapa Zagrożeń Bezpieczeństwa) is a key tool in the process of managing public security in Poland. Its purpose is to identify and present the scale and type of threats that can affect the sense of security of local communities. The map is unique because of its multidimensional data structure. The information on which it is based comes from three main sources. The first is data collected in police information systems, which provide objective statistical data on crimes and offenses. The second source is information obtained directly from citizens, representatives of local government, non-government organizations, etc., which is collected during direct contacts and during public debates on public security. The third source is information obtained from Internet users through a special information sharing platform [44], as seen in Fig. 1.
The National Security Threat Map considers both selected categories of crimes and offenses, as well as threats that, in the subjective perception of residents, negatively affect their sense of security. Through this tool, it is possible not only to understand what threats are most relevant to the community, but also what the public's security expectations are. In addition, this map is useful in the decision-making process regarding the allocation of equipment and personnel resources for security services. It can help inform decisions on the establishment of police stations and precincts, allowing resources to be directed to where they are most needed [44].

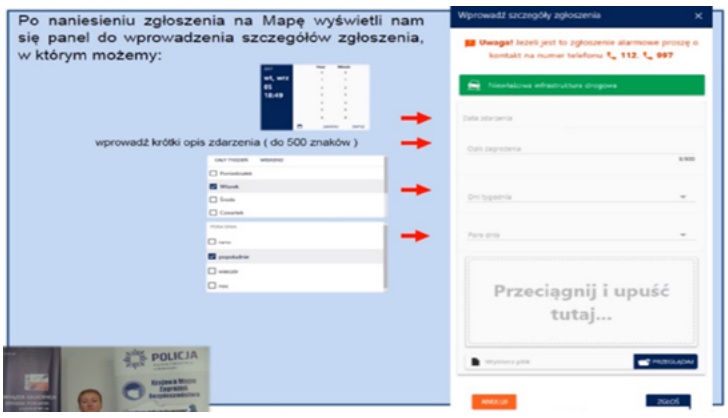

Fig. 1. Adding a notification to the National Safety Threats Map Source: Training materials of the Polish Police Forces [44]

The National Security Threat Map considers both selected categories of crimes and offenses, as well as threats that, in the subjective perception of residents, negatively affect their sense of security. Through this tool, it is possible not only to understand what threats are most relevant to the community, but also what the public's security expectations are. In addition, this map is useful in the decision-making process regarding the allocation of equipment and personnel resources for security services. It can help inform decisions on the establishment of police stations and precincts, allowing resources to be directed to where they are most needed [44].

4. DATA OBTAINING PROCESS

In his paper [45] Badurowicz in 2022 has proven that extracting data using web scraping methods and then building machine learning models based on them can be a very effective approach. They built a program to analyze the content of a software forum based on automatic machine learning models and achieved a 95% success rate in detecting untagged source code. Another study, based on posts obtained from Twitter during Hurricane Florence in September 2018, explored the possibilities of providing a multifaceted and detailed picture of events taking place in the affected areas [46]. To obtain the data, the study considered two criteria for data selecting. As a priority, the assumption was the data should be easily obtained without additional third-party consent. The commonly used API greatly simplifies the data collecting process but limits the tool's independence and limits its universality. The selection of data sources, based on the latest research found on the Ortiz-Ospina & Roser subject [47] considered the most popular social media such as: Facebook, YouTube, WhatsApp, Instagram, WeChat, TikTok, Telegram and Twitter. Research demonstrated that in Poland Facebook, YouTube, and Instagram are being used by users most frequently [48, 49].

4.1. Threats detectable on social media

A Storey in his study in 1995 described the language people use during a threat as expressions that leave a negative effect, such as injury, material loss or stress [50]. Although the scale of potential threats is unlimited, not all of them are relevant in the context of the local community and not all of them are possible to detect. For example, so-called hate crime, despite being common and easily detectable phenomenon, has more negative effects on victims' mental health than direct impact on the security of the local community [51]. Some local phenomena that may require a quick response and occur relatively frequently based on study [52], are shown in Fig. 2. Threats were grouped, and the resulting collections were used as labels in further research.

Fig. 2. Tree diagram of threat groups

4.2. Applied selection criteria for selection of data sources

The first adopted criterion is accessibility. In order to effectively analyze actively influencing data and expand the scope of the source of analysis, it is required that this data be easily accessible. Sources with difficult access to data and those whose analysis is significantly difficult have been rejected. The second criterion that guided the selection was popularity. Sources that are not popular in the studied region and those that do not provide adequate information on the topic of threats were rejected. The popularity of social media services may be different in other countries, yet services such as Facebook, YouTube, Instagram and WeChat are the most popular worldwide [47].

Facebook was chosen as the source that best fits the objectives. Facebook groups are a popular source of information in every region of Poland, including Subcarpathian Province. The groups bring together a significant portion of local communities and actively report possible threats. The study was reduced to three threats that must be dealt with as quickly as possible to reduce their impact. Threats that occur relatively infrequently were discarded in variant with human expert. Tool ChatGPT was able to vote binary. In this study, the groups shown in Fig. 2 were focused on.

4.3. Testing the process of data acquisition and machine learning techniques

In order to obtain the data, a tool was built to copy user posts from local Facebook groups. The tool requires the completion of a list of groups to be periodically analyzed. The application is based on Django's Python library and is a cloud-based solution. The application does not analyze any information about the authors of the entries, only the content of the entries.

The data collection application was developed using the Django framework, which is dedicated to web solutions. The Selenium tool was used to automate browser processes, which allows interaction with dynamic websites. As part of the project, the application was configured to run with the Chrome browser. To optimize the data collection process, the Celery library was used to queue and manage tasks. Due to the various anti-bot mechanisms used by most websites, the program was limited to downloading data from one Facebook group at a time, using a one-by-one loop approach.

In order to start periodically scanning Facebook groups using the app, the administrator needs to provide Facebook account login credentials and a list of groups to be periodically checked. In addition, settings such as the intervals between browser actions or the number of times the page is scrolled can be adjusted. Before launching the application, the user must independently join these groups on the Facebook platform. Then, thanks to the Celery library, it is possible to periodically launch the browser and retrieve data from each group.

Due to the dynamic nature of the Facebook page, a special approach to the scraping technique was required. For each group using JavaScript, the page was scrolled down several times to load the content of the posts, then the „See more" button was triggered for each post, and the corresponding page elements were set to „Active" to download the dynamically generated data. To minimize the duplicate posts in the database, a duplication check mechanism is used. When writing to the database, checksums are calculated based on the content of the entry and the date when the entry was added.

If a given counter sum exists in the database, then the entry is skipped. The entire process is summarized in Fig. 3. The dataset collected is available in data availability statement at the end of the paper.

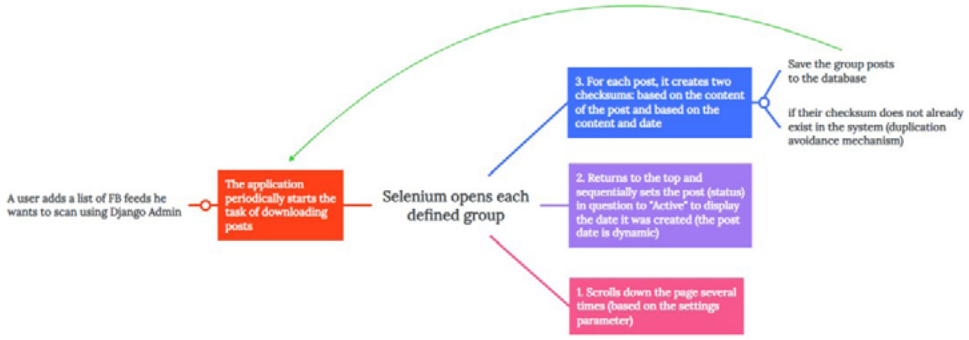

Fig. 3. Mechanism of the web scraper application functioning

5. RESULTS AND DISCUSSION

The data was analyzed in three variants. The initial variant (I) used ChatGPT to estimate categories (labels) in a binary form. Then the expert reanalyzed the data and made adjustments (variant II) and manually expanded the labels from binary to 3 groups (variant III). The data analysis process for variants II and III is shown in Fig. 4. Additionally, variant I provided a basis for accelerating the Expert's work. In order to speed up the research and labeling process, the entries were evaluated using the ChatGPT tool. A query was performed for each entry using the prompt featured in Listing 1.

Fig. 4. Data analysis process for variant II

„Is post xxx a threat. Answer yes or no."
(Original polish language: "Czy post xxx jest zagrożeniem. Odpowiedź tak lub nie".)

Listing. 1. Data analysis process for variant II.

Tab. 1. Unique responses obtained through the API of the ChatGPT service

Unique (Original)	Unique (ENG)	Occurrence
Nie.	No.	379
Nie	No	9
Tak.	Yes.	24
nie	no	9

Tab. 2. Aggregated responses of the ChatGPT service

Aggregated responses of the ChatGPT service	Occurrence
Positive (threat detected)	24
Negative (threat undetected)	397

Tab. 3. The quality of the labels acquired using ChatGPT

Samples	Threats detected by ChatGPT	Confirmed samples by Expert	Additional threats identified by the Expert (false negative)	Rejected threats by the Expert (false positive)
421	24	389	19	4

Tab. 4. Unique responses obtained through the API of the ChatGPT service

Unique (Original)	Unique (ENG)	Occurrence
Utrudnienia	Difficulties	24
Przestępczość	Crime	11
Naturalne	Natural	4

The service, through the API, returned the responses contained in Tab 1. Labels grouped to negative and positive. As a result of the above operations, it was estimated that 6% of posts were evaluated as a threat. The prepared labels served as an aid to the expert's evaluation of the data and to the preliminary analysis of effectiveness presented in section 5.1. Based on the data in Tab 2., the expert made adjustments and manually grouped the data by assigning appropriate labels. The quality of the extracted labels is included in Tab 3. The distribution of the data labeled by Expert is included in Tab 4.

5.1. Data analysis with machine learning methods

From the data set, all punctuation marks, commas, stop words have been removed, and all words have been changed to lowercase. The content was divided into individual units called tokens. Since machine learning models are not able to operate directly on text, it was necessary to convert words into numbers as their representation. Once the data was prepared, the ML model was used to test the ability to detect the assumed threats contained in user content. For three variants ML model was developed and tested.

Tab. 5. Results for ChatGPT labeled dataset (variant I)

Model name	Accuracy [%]	Model name	Accuracy [%]
AdaBoostClassifier	0.92	LinearDiscriminantAnalysis	0.66
BaggingClassifier	0.95	LinearSVC	0.92
BernoulliNB	0.90	LogisticRegression	0.94
CalibratedClassifierCV	0.94	NearestCentroid	0.94
DecisionTreeClassifier	0.92	PassiveAggressiveClassifier	0.92
DummyClassifier	0.94	Perceptron	0.91
ExtraTreeClassifier	0.93	QuadraticDiscriminantA-nalysis	0.64
ExtraTreesClassifier	0.94	RandomForestClassifier	0.94
GaussianNB	0.92	RidgeClassifier	0.94
KNeighborsClassifier	0.94	RidgeClassifierCV	0.94
LGBMClassifier	0.94	SGDClassifier	0.94
LabelPropagation	0.94	SVC	0.94
LabelSpreading	0.94		

Tab. 6. Results for expert labeled dataset (binary labels, variant II)

Model name	Accuracy [%]	Model name	Accuracy [%]
AdaBoostClassifier	0.93	LinearDiscriminantAnalysis	0.92
BaggingClassifier	0.93	LinearSVC	0.90
BernoulliNB	0.88	LogisticRegression	0.92
CalibratedClassifierCV	0.92	NearestCentroid	0.92
DecisionTreeClassifier	0.93	PassiveAggressiveClassifier	0.92
DummyClassifier	0.92	Perceptron	0.95
ExtraTreeClassifier	0.94	QuadraticDiscriminantA-nalysis	0.92
ExtraTreesClassifier	0.93	RandomForestClassifier	0.92
GaussianNB	0.94	RidgeClassifier	0.92
KNeighborsClassifier	0.92	RidgeClassifierCV	0.92
LGBMClassifier	0.92	SGDClassifier	0.91
LabelPropagation	0.92	SVC	0.92
LabelSpreading	0.92		

Tab. 7. Results for expert labeled dataset (tree labels, variant III)

Model name	Accuracy [%]	Model name	Accuracy [%]
AdaBoostClassifier	0.92	LinearDiscriminantAnalysis	0.67
BaggingClassifier	0.92	LinearSVC	0.90
BernoulliNB	0.88	LogisticRegression	0.92
CalibratedClassifierC	0.92	NearestCentroid	0.92
DecisionTreeClassifier	0.96	PassiveAggressiveClassifier	0.91
DummyClassifier	0.92	Perceptron	0.95
ExtraTreeClassifier	0.92	QuadraticDiscriminantAnalysis	0.04
ExtraTreesClassifier	0.92	RandomForestClassifier	0.92
GaussianNB	0.92	RidgeClassifier	0.92
KNeighborsClassifier	0.92	RidgeClassifierCV	0.92
LGBMClassifier	0.92	SGDClassifier	0.91
LabelPropagation	0.92	SVC	0.92
LabelSpreading	0.92		

ISSN 2082-7016; eISSN 2450-5552 2023, nr 17

For the three variants, ML model learning and split testing were performed. The results for variant I are shown in Tab 5, for variant II in Tab 6, and for variant III in Tab 7. and visualized on Fig. 5. and Fig. 6. Additionally for the algorithm with highest accuracy (Decision Trees and variant III), cross-validation was additionally checked: for 10-kfold it was 0.94% and for 5-kfold – 0.93%.

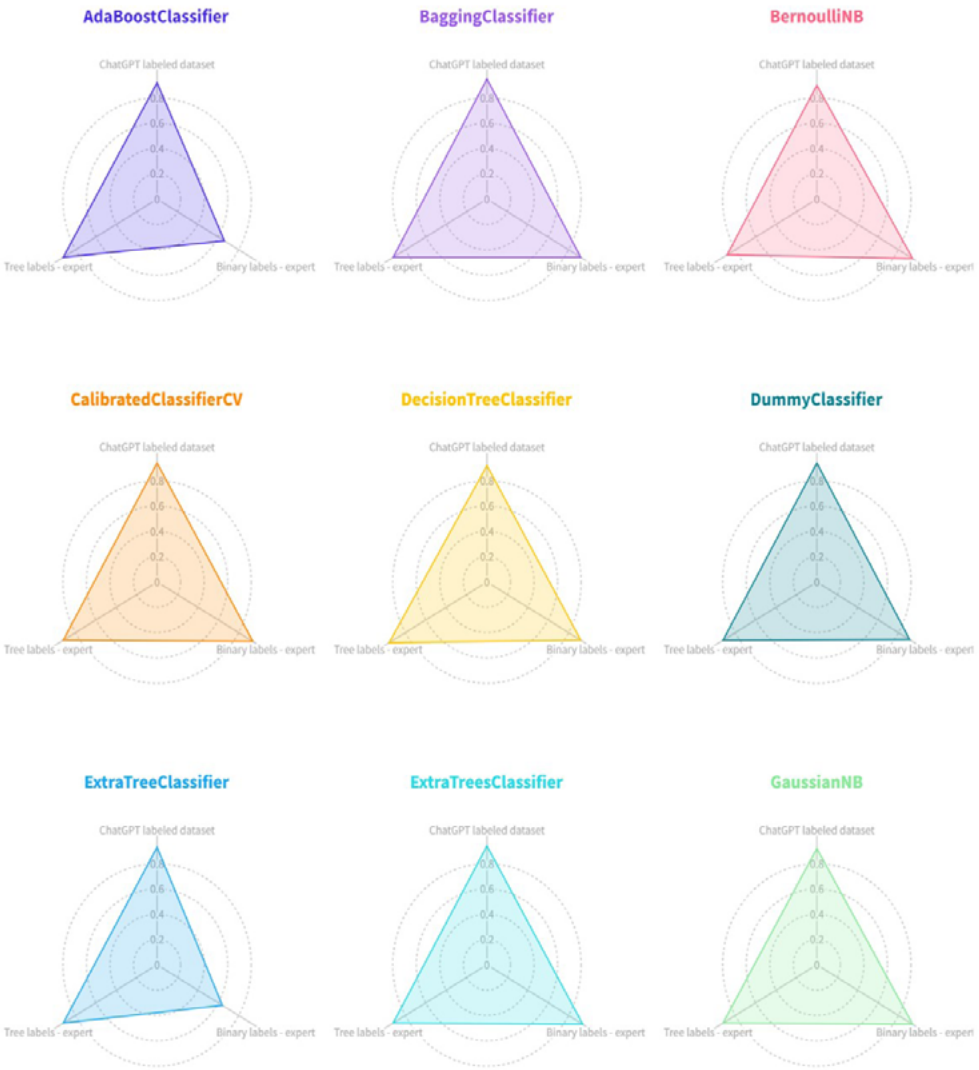

Fig. 5. Results visualization – part I

Fig. 6. Results visualization – part II

5.2. Discussion

When comparing the efficiency of ChatGPT's initial categorization (Variant I) with expert-verified data (Variants II and III), it is interesting to observe that the former exhibited a slightly better performance in terms of accuracy across several machine learning models. This could be attributed to the robustness of the ChatGPT model in understanding and classifying text-based data. However, it's crucial to remember that the model still made some false positive and false negative errors, indicating areas for potential refinement. The most striking performance difference was observed in the Decision Tree Classifier model when binary labels were expanded to three categories (Variant III), achieving a classification accuracy of 96%. The notable increase in accuracy might be due to the addition of more categorical detail, which allows the model to make more nuanced decisions in classifying the data. This finding suggests that fine-grained categorization of threat labels might improve threat detection accuracy. Cross-validation of the top-performing Decision Tree Classifier further demonstrated its reliability, with consistently high accuracy across 5-kfold and 10-kfold cross-validations. However, while the model performed impressively, a significant number of threats remained undetected. This could be due to the small size of the data set and the particularly low number of samples identified as threats, reinforcing the need for larger data sets for model training and evaluation.

The study reaffirms the potential of AI and machine learning in threat detection, which could have significant implications for various fields, including cybersecurity [53] and content moderation. It also highlights the necessity for continuous improvement and refinement in the tools used, including hybrid approaches that combine AI and expert analysis. Future studies should explore these considerations, leveraging larger and more diverse data sets to improve model performance in threat detection.

6. CONCLUSIONS

In the paper, the Subcarpathian Province was selected as the subject of the study in order to initially test the validity of the proposed approach. The proposed method for detecting threats and crisis events can be applied to any administrative services where particular social media apps are popular. The ChatGPT tool proved to be a good and fast tool for analyzing the studied problem, but it has its drawbacks. The finished ML model will be much cheaper than ChatGPT tool in the long run. Despite the drawbacks, pre-labeling with this tool greatly speeds up the expert's work. The tool could be one of the security factors of administrative centers. It reduces the costs associated with daily manual analysis by humans. Appropriate parameterization of the application can show more vulnerable areas, report only the most dangerous events or cases where many users reported a threat in a short period of time. The proposed solution does not exhaust the topic in any aspect and requires further research work and functional testing. There are many possible developments of the proposed approach, these include increasing the number of threats under investigation, increasing the number of data sources and better matching them. Adding other media for analysis such as images and video is worth considering. Improving the effectiveness of the solution could be achieved by analyzing the number of detections of similar threats per unit of time, or the number of negative user reactions to an alert.

Data availability statement
The data that support the findings of this study are openly available in GitHub Repository at *www.github.com/Vitz/threats_dataset*

Author Contributions
All authors declare equal contribution to this research paper.

Acknowledgments
This project was financed by funds for the development of the research potential of the dissertation „Technical and Telecommunications Informatics" from the funds of the Department of Complex Systems in the Rzeszow University of Technology.

Conflicts of Interest
The authors declare there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

REFERENCES

[1] Aillerie, K., & McNicol, S. (2018). Are social networking sites information sources? Informational purposes of high-school students in using SNSs. Journal of Librarianship and Information Science, 50(1), 103-114.
[2] Giles, H., & Ogay, T. (2007). Communication accommodation theory.
[3] Sinta, I., Ilham, R. N., ND, M. A., Subhan, M., & Usman, A. (2022). UTILIZATION OF DIGITAL MEDIA IN MARKETING GAYO ARABICA COFFEE. IRPITAGE JOURNAL, 2(3), 103-108.
[4] Audrezet, A., De Kerviler, G., & Moulard, J. G. (2020). Authenticity under threat: When social media influencers need to go beyond self-presentation. Journal of business research, 117, 557-569.,
[5] Boulianne, S., Lalancette, M., & Ilkiw, D. (2020). „ School Strike 4 Climate": Social Media and the International Youth Protest on Climate Change. Media and Communication, 8(2), 208-218.
[6] Abbas, J., Wang, D., Su, Z., & Ziapour, A. (2021). The role of social media in the advent of COVID-19 pandemic: crisis management, mental health challenges and implications. Risk management and healthcare policy, 1917-1932.
[7] Ferrara, E., Cresci, S., & Luceri, L. (2020). Misinformation, manipulation, and abuse on social media in the era of COVID-19. Journal of Computational Social Science, 3, 271-277.
[8] Rahmayani, O., Ardi, S., & Nofrialdi, R. (2022). The Effect of Utilization of Social Media Instagram@ Nanarfshop on Buying Interest of Fisipol Students University Ekasakti Padang. Journal of Law, Politic and Humanities, 2(2), 85-94.
[9] Dubbelink, S. I., Herrando, C., & Constantinides, E. (2021). Social media marketing as a branding strategy in extraordinary times: Lessons from the COVID-19 pandemic. Sustainability, 13(18), 10310.
[10] Pan, B., & Crotts, J. C. (2016). Theoretical models of social media, marketing implications, and future research directions. In Social Media in Travel, Tourism and Hospitality (pp. 73-86). Routledge.
[11] LoBue, V., Matthews, K., Harvey, T., & Stark, S. L. (2014). What accounts for the rapid detection of threat? Evidence for an advantage in perceptual and behavioral responding from eye movements. Emotion, 14(4), 816.
[12] Bessarab, A., Mitchuk, O., Baranetska, A., Kodatska, N., Kvasnytsia, O., & Mykytiv, G. (2021). Social Networks as a Phenomenon of the Information Society. Journal of Optimization in Industrial Engineering, 14(Special Issue), 17-24. doi: 10.22094/joie.2020.677811
[13] Rost, M., Barkhuus, L., Cramer, H., & Brown, B. (2013, February). Representation and communication: Challenges in interpreting large social media datasets. In Proceedings of the 2013 conference on Computer supported cooperative work (pp. 357-362).
[14] Goban-Klas, T. (2004). Media i komunikowanie masowe: Teorie i analizy prasy, radia, telewizji i Internetu. Wydawnictwo Naukowe PWN SA.
[15] Kaznowski, D. (2008). Nowy marketing. VFP Communications.
[16] Koczkodaj, W. W., Kowalczyk, A., Mazurek, M., Pedrycz, W., Redlarski, G., Rogalska E., Strzalka D., Szymanska A., Wilinski A., Xue, O. S. (2023). Peer assessment as a method for measuring harmful internet use. MethodsX, 11, 102249. doi:10.1016/j.mex.2023.102249
[17] TEDx Talk - Bailey Parnell - Is Social Media Hurting Your Mental Health? [Video]. (n.d.). Retrieved from *www.youtube.com/watch?v=Czg_9C7gw0o*
[18] Steers, M. L. N., Wickham, R. E., & Acitelli, L. K. (2014). Seeing everyone else's highlight reels: How Facebook usage is linked to depressive symptoms. Journal of Social and Clinical Psychology, 33(8), 701-731.
[19] Magdol, L., & Bessel, D. R. (2003). Social capital, social currency, and portable assets: The impact of residential mobility on exchanges of social support. Personal Relationships, 10(2), 149-170.
[20] Jupowicz-Ginalska, A., Jasiewicz, J., Kisilowska, M., Baran, T., & Wysocki, A. (2018). FOMO. Polacy a lęk przed odłączeniem–raport z badań [FOMO. Poles and the Fear of Missing Out–A Research Report]. Warszawa: Wydział Dziennikarstwa Informacji i Bibliologii UW.

[21] Vogels, E. A. (2021). The state of online harassment. Pew Research Center, 13, 625.
[22] Solove, D. J. (2007). I've got nothing to hide and other misunderstandings of privacy. San Diego L. Rev., 44, 745.
[23] Solove, D. J. (2011). Why privacy matters even if you have ‚nothing to hide'. Chronicle of Higher Education, 15.
[24] Anthes, G. (2014). Data brokers are watching you.
[25] Crain, M. (2018). The limits of transparency: Data brokers and commodification. New media & society, 20(1), 88-104.
[26] Kadam, A. B., & Atre, S. R. (04 2020). Negative impact of social media panic during the COVID-19 outbreak in India. Journal of Travel Medicine, 27(3), taaa057. doi:10.1093/jtm/taaa057
[27] Lelisho, M. E., Pandey, D., Alemu, B. D., Pandey, B. K., & Tareke, S. A. (2023). The Negative Impact of Social Media during COVID-19 Pandemic. Trends in Psychology, 31(1), 123–142. doi:10.1007/s43076-022-00192-5
[28] Kilic, S. O. (2023). Turkey earthquake: The media did not hold the government to account. Middle East Eye. Retrieved from *www.middleeasteye.net/opinion/turkey-earthquake-media-not-hold-government-account*
[29] Northeastern University News. (2023, February 8). How social media played a crucial role during the Turkey earthquake. Retrieved from *www.news.northeastern.edu/2023/02/08/social-media-turkey-earthquake/*
[30] The Guardian. (2023, February 14). How misinformation hampered relief efforts during the Turkey-Syria earthquake. Retrieved from *www.theguardian.com/commentisfree/2023/feb/14/turkey-syria-earthquake-misinformation-relief-efforts-turkey*
[31] Cox, S. R., & Atkinson, K. (2018). Social media and the supply chain: Improving risk detection, risk management, and disruption recovery. SAIS 2018 Proceedings, 8. Retrieved from *www.aisel.aisnet.org/sais2018/8*
[32] Alexander, D. (2014). Social media in disaster risk reduction and crisis management. Science and Engineering Ethics, 20(3), 717-733.
[33] Chen, X., Vorvoreanu, M., & Madhavan, K. (2014). Mining social media data for understanding students' learning experiences. IEEE Transactions on learning technologies, 7(3), 246-259.
[34] López-Vizcaíno, M. F., Nóvoa, F. J., Carneiro, V., & Cacheda, F. (2021). Early detection of cyberbullying on social media networks. Future Generation Computer Systems, 118, 219–229. doi:10.1016/j.future.2021.01.006
[35] Wang, T., Wang, T., Cheng, G. Y., & Yue, Y. (2012). A model of Internet public opinion pre-warning based on fuzzy comprehensive evaluation. J. Intell., 6, 47-52.
[36] Cao, H. J., Li, M., & Hou, T. T. (2019). An empirical study of Internet public opinion early warning based on elders-related emergencies: Take the event of ‚elderly square dancers forcibly occupy the basketball court' as an example. J. Chongqing Univ. Posts Telecommun. (Social Sci. Ed.), 1, 58-66.
[37] Mazurek, J., Perzina, R., Strzałka, D., Kowal, B., & Kuraś, P. (2021). A numerical comparison of iterative algorithms for inconsistency reduction in pairwise comparisons. IEEE Access, 9, 62553-62561.
[38] Koczkodaj, W. W. (1993). A new definition of consistency of pairwise comparisons. Mathematical and computer modelling, 18(7), 79-84.
[39] Zhao, Q., & Pan, Y. T. (2020). Research on early warning of public opinion based on BP neural network. Statist. Appl., 9(2), 224-236.
[40] RCB, Biuletyn analityczny Rządowego Centrum Bezpieczeństwa nr 32-33, Rządowe Centrum Bezpieczeństwa.
[41] Council of Ministers of Poland, Rozporządzenie Rady Ministrów z dnia 7 grudnia 2018 r. w sprawie współpracy dyrektora Rządowego Centrum Bezpieczeństwa z operatorem ruchomej publicznej sieci telekomunikacyjnej w celu powiadamiania użytkowników końcowych o zagrożeniu.
[42] RCB, Raport dobowy Rządowego Centrum Bezpieczeństwa, 09.06.2023.
[43] Nestorenko T, Ostenda A., 2021, Selected aspects of digital society development, Katowice, Publishing House of University of Technology.

[44] Krajowa Mapa Zagrożeń Bezpieczeństwa (2023). Retrieved from *www.mapy.geoportal.gov.pl/iMapLite/KMZBPublic.html*
[45] Badurowicz, M. (2022). Detection of source code in internet texts using automatically generated machine learning models. Applied Computer Science, 18(1), 89-98. *www.doi.org/10.23743/acs-2022-07*
[46] Kersten, J., & Klan, F. (2020). What happens where during disasters? A Workflow for the multifaceted characterization of crisis events based on Twitter data. Journal of Contingencies and Crisis Management, 28(3), 262–280. doi:10.1111/1468-5973.12321
[47] Ortiz-Ospina, E., & Roser, M. (2023). The rise of social media. Our world in data.
[48] Król, K., & Zdonek, D. (2021). Most Often Motivated by Social Media: The Who, the What, and the How Much—Experience from Poland. Sustainability, 13(20), 11193. MDPI AG. Retrieved from *www.dx.doi.org/10.3390/su132011193*
[49] Werenowska, A., & Rzepka, M. (2020). The Role of Social Media in Generation Y Travel Decision-Making Process (Case Study in Poland). Information, 11(8), 396. MDPI AG. Retrieved from *www.dx.doi.org/10.3390/info11080396*
[50] Storey, K. (1995). The language of threats. International Journal of Speech, Language and the Law, 2(1), 74-80.
[51] Müller, K., & Schwarz, C. (10 2020). Fanning the Flames of Hate: Social Media and Hate Crime. Journal of the European Economic Association, 19(4), 2131–2167. doi:10.1093/jeea/jvaa045
[52] Kahn, D. T., Björklund, F., & Hirschberger, G. (2022). The intent and extent of collective threats: A data-driven conceptualization of collective threats and their relation to political preferences. Journal of Experimental Psychology: General, 151(5), 1178.
[53] Matejkowski, D., & Szmyd, P. (2023). Online identity theft detection and prevention methods. Advances in Web Development Journal, 1(1), 12. *www.doi.org/10.5281/zenodo.10051152*

WYKRYWANIE ZAGROŻEŃ I ZDARZEŃ KRYZYSOWYCH Z WYKORZYSTANIEM TECHNIK UCZENIA MASZYNOWEGO W OPARCIU O DANE Z MEDIÓW SPOŁECZNOŚCIOWYCH

STRESZCZENIE

W artykule autorzy przedstawiają wyniki prac nad oprogramowaniem web scrapingowym pozwalającym na zautomatyzowaną klasyfikację zagrożeń i wykrywanie zdarzeń kryzysowych. W celu poprawy bezpieczeństwa i komfortu życia ludzi przeprowadzono analizę szybkiego wykrywania zagrożeń z wykorzystaniem nowoczesnego kanału informacyjnego jakim są media społecznościowe. W tym celu dokonano przeglądu popularnych w badanym regionie serwisów społecznościowych i wybrano odpowiednie, kierując się kryteriami dostępności i popularności. Zebrano i przeanalizowano około 300 unikalnych postów z lokalnych grup miast i innych ośrodków administracyjnych. Decyzja o tym, który wpis został sklasyfikowany jako zagrożenie, została określona przy użyciu narzędzia ChatGpt oraz przy udziale osoby (eksperta). Oba warianty zostały przetestowane przy użyciu metod uczenia maszynowego (ML). Dodatkowo, w artykule sprawdzono, czy narzędzie ChatGpt będzie skuteczne w wykrywaniu domniemanych zdarzeń i porównano to rozwiązanie z klasycznym podejściem ML, gdzie dane uczące etykietowano przy udziale ekspretra.

SŁOWA KLUCZOWE

media społecznościowe, wykrywanie zagrożeń, web scraping, chatgpt, uczenie maszynowe

Artykuł udostępniony na licencjach Creative Commons/ Article distributed under the terms of Creative Commons licenses: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0). License available: www.creativecommons.org/licenses/by-nc-sa/4.0/