REPORT

Firearms violence incident monitor – Methodological report

flemish peace institute

Colophon

Firearms violence incident monitor – Methodological report

ISBN: 9789464074260

Authors: Diederik Cops, Astrid De Schutter & Tom De Smedt

Editing: John Linnegar

Publisher: Nils Duquet, Leuvenseweg 86, 1000 Brussel

CONTENTS

Introduction 4

Measuring the scope of firearms violence in the EU 4

1 Firearms incident monitor using Artificial Intelligence 6

1.1 Selecting relevant media sources and identifying relevant articles on gun violence incidents 9

1.2 Clustering different articles on the same incident 17

1.3 Extracting relevant information on gun violence incidents from media articles 18

2 Concluding remarks, lessons learned 21

Introduction: measuring the scope of firearms violence in the EU

In its Action Plan on Firearms Trafficking 2020-2025, the European Commission (EC) notes that the fight against illicit access to firearms must be a cross-thematic security priority for the European Union (EU), its Member States and its partners. As the EC indicates, illicit access to firearms increases the risk of crime-related and domestic violence that may escalate into murder. Illegal access to firearms also facilitates other forms of criminal activity. In order to respond adequately to the threat posed by the illicit trafficking of firearms, the Action Plan contains four priorities. Under Priority 2 (Building a better intelligence picture) the plan notes that, in cooperation with Europol, the EC will explore the feasibility of rolling out a tool to track firearms-related incidents in real time and developing a permanently up-to-date picture [Action 2.6 – KPI 10].[1]

Comprehensive and comparable data on firearms incidents in the EU are currently scarce and have to be interpreted with caution. With this in mind, project INSIGHT (International Network Studying Incidents with Guns: Harm & Trafficking) was developed to meet the need to build a better intelligence picture by creating an online knowledge platform that brings together all information on firearms violence and firearms trafficking in the 27 EU Member States. More specifically, the project contributes to strengthening the intelligence picture of illicit firearms trafficking and gun violence in the EU by developing a tool to track firearms-related incidents and seizures in real time. The project achieves this by pooling existing aggregated data from government agencies and by delivering state-of-the-art knowledge products on these security phenomena. The project received funding via the Internal Security Fund (ISF) of the EU and ran from January 2022 until December 2023.

The project consortium comprises

- four research partners with vast policy-relevant research experience in illicit firearms trafficking and gun violence (Flemish Peace Institute, Small Arms Survey, SIPRI and Leiden University);
- a research partner that specialises in innovative artificial intelligence (AI) technology to generate disaggregated data (Textgain);
- three operational law-enforcement partners (Europol, European Firearms Experts and the Dutch National Police); and
- an operational partner with experience in developing a real-time tracking tool on gun violence in the Western Balkans (SEESAC).

By generating data-driven intelligence about firearms-related incidents and seizures, this project aimed to:

- develop more robust, adequate and effective policies and operational interventions;
- raise awareness about the advantages of better data collection and analysis;
- promote best practices for defragmenting relevant information and intelligence; and
- strengthen both national and international information-sharing between law-enforcement agencies.

This methodological report provides a description of the development of the online monitor as part of Project INSIGHT. The monitor aims to generate an almost real-time and automated identification of firearms incidents in all the EU Member States that is easily accessible to all stakeholders. Publicly available media articles are used to identify these incidents. The monitor implements and deploys different methods of AI, such as machine learning (ML) and large language models (LLMs), to automate the process of identifying, assessing, clustering and analysing media articles on firearms violence and firearms seizures in the EU.

In consequence of our building on expert knowledge of firearms and firearms violence from the researchers and research institutes in the consortium and the extensive training material gathered by the research consortium, the different AI methods have been trained (and will continue to be trained) to identify, cluster and analyse media articles automatically for the duration of the project (from 1 January 2022 until December 2023).

In the remainder of this report, the different stages of generating the monitor of incidents of firearms violence and how they have been developed are described in depth. The first section of this methodological report elaborates on the subsequent steps and the AI methods on which the online monitor is built. This description is intended to provide a better understanding of the methods that underpin the monitor, how the information included in the monitor is identified, assessed and analysed, and what steps have been taken to enhance the reliability and validity of the data included in the monitor. In doing so, we trust that this report will also lead to a better understanding of the limitations of and the opportunities for using media sources to identify instances of gun violence in the EU and to extract relevant information on these cases as provided by media articles. This report ends with some final remarks and reflections on the comprehensiveness of the monitor and its relationship to other datasets that pertain to gun violence.

1 Firearms incident monitor using Artificial Intelligence

Previous studies have revealed the substantial lack of reliable, comprehensive and detailed data on firearms violence in the EU.[2] As a consequence, building a comprehensive picture of this phenomenon – which is crucial to devising an effective policy that deals with and resolves this problem – is virtually impossible. Additional sources besides official police and judicial data are therefore needed to help gain a better overview and understanding of (developments in) the scope of incidents of firearm violence in the EU and the context in which they occur.

The use of newspaper articles as a source of information on (violent) crime is common in scientific research, government-issued reports and also in-depth investigations conducted by journalists.

The use of media reports has proved to be useful in previous studies for mapping violence-related phenomena such as gang homicides in Canada,[3] the illegal use of hand grenades in the Netherlands[4] and murder-suicides.[5] While these studies used newspaper articles to establish the nature and scope of firearms violence retrospectively, several initiatives have been developed to collect and publish information on such incidents based on media reports in a more real-time manner. In the European context, the Armed Violence Monitoring Platform by SEESAC is a notable example: it registers firearm-related incidents in south-east Europe by monitoring the local media and relevant reports by public institutions.[6] Other channels, such as the website gunviolencearchive.org, also collect reports on gun violence and crime incidents using public media sources in the United States.

Media reports on incidents of firearm violence have several advantages. These are often more detailed than official police press releases, because they are updated as more information becomes available (to a certain point) and often contain information that is not found in other sources – for example, the background of the victim or the perpetrators gained through interviews with family members and the specific context of an incident. In particular, media articles provide information about the perpetrators, their presumed motives, the victims, the type of violence and the spatial–temporal characteristics of events.

However, it must be made clear from the outset that **publicly available media sources by no means make it possible to form a comprehensive understanding of gun violence incidents or seizures of firearms**. One reason is that not all incidents in which a firearm was used are reported in the media. Another is that media attention on firearm incidents will be directed disproportionately to more spectacular events, resulting in the under-reporting of specific types of crime, in particular non-violent crime such as threats with firearms or violence that takes place in the private sphere, such as domestic violence involving firearms. Moreover, the information included in media reports will not necessarily always be reliable, as the information obtained might not always be correct or some information may not be sufficiently relevant or interesting to be included. These factors can therefore make it difficult to integrate such information into an analysis.

The drawbacks of a media analysis can therefore be fourfold:

1. the information in the newspapers is usually not verified against official sources and so it tends to be less reliable;
2. media articles are also published most often before a final judicial decision is taken in court, which means that the exact legal classification of an incident can be unclear (official decisions on cases are rarely reported in the media);
3. media articles mostly lack more precise information about firearms and, when it is indeed available, it is mostly unreliable;
4. media attention is not given to events objectively, but often it is based on what a society considers to be of interest.[7]

Finally, and, as previously mentioned, the publication of media articles on firearms incidents tends to lag somewhat behind the occurrence of an incident. In most cases, an article is published only the day after an incident has occurred, which results in a short time lag. This suggests that a monitor such as ours which uses media articles as a source of information is not completely real-time.

This section describes the three stages or steps which constitute the online monitor of incidents of firearms violence that has been developed:

1. Identify and select relevant media articles on incidents of firearms violence in all EU Member States (1.1).
2. Cluster the different media articles that cover the same incident of firearms violence (1.2).
3. Extract from the media articles the relevant contextual information regarding an incident involving firearms violence (1.3).

The different steps taken to increase the **reliability** and **validity** of the information included in the real-time monitor are also be discussed in this report. 'Reliability' refers to the extent to which a specific research method is consistent in its measurement, a method being reliable if the same result can be achieved consistently by using the same method under the same circumstances. In this context, reliability refers to the extent to which articles relating to firearms incidents are correctly identified and to which articles that do not relate to firearms incidents are correctly discarded. In ML this reliability can be statistically evaluated by precision and recall. For instance, the overall precision and recall of AI in this context has found to be 90%. It is slightly better at discarding unrelated articles (93%) than it is at identifying related articles (87%) (see further in the report). However, an in-depth analysis shows that, in the case of about 10% of articles that are not clear cut, human beings also differ on whether articles are related or not; sometimes AI does even better at discarding outdated events or identifying reports about court cases, etc.

Validity, which refers to the extent to which a method used actually measures what it aims to measure, has two aspects: internal and external validity. **Internal validity** refers to the extent to which the articles identified as firearms incidents are effectively articles that fall within the scope of what are defined as firearms incidents. **External validity** here refers to the extent to which all incidents that are reported on in media sources are being identified by the monitor effectively.

The monitor was developed using Textgain's Grasp software (https://github.com/textgain/grasp).

Figure 1: Schematic overview – flowchart of the different steps taken in monitoring firearms violence incidents

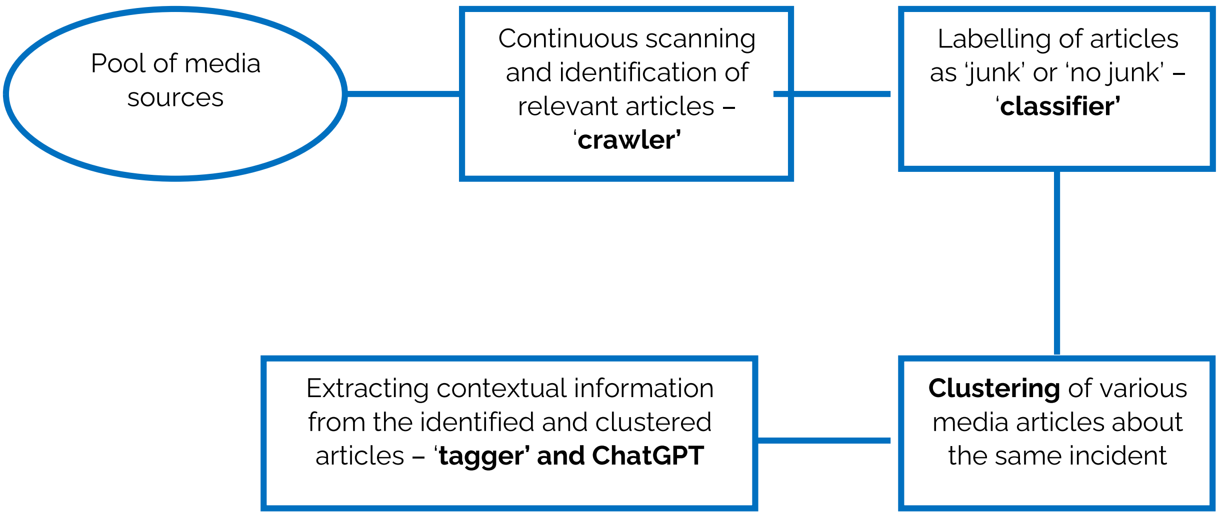

Figure 1 provides a schematic overview of the different steps on which the firearms incidents monitor is built. These sequential steps and their development throughout a project are explained in the remainder of this report.

1.1 Selecting relevant media sources and identifying relevant articles on gun violence incidents

This subsection describes the different parts that result in the identification of relevant media articles on firearms violence. It also describes the human steps that were taken to select the relevant media sources (1.1.1) in order to develop the list of key words that help to identify relevant articles and to cause an instrument to label the identified articles as either ‘junk’ or ‘no junk’ (1.1.2). These two steps were taken and developed simultaneously, because both a list of media sources and a list of keywords or phrases by which to identify relevant media articles are indispensable.

1.1.1 Selection of newspapers and of potentially relevant articles

A crucial first element was the development of a **list of media sources** that would be used to identify those media articles that covered gun violence. Based on the previous experience and expertise in the research team regarding the identification of such articles, a combination of national and regional media sources was sought for each EU Member State. This list was composed manually by the research partners of Project INSIGHT and was expanded and fine-tuned incrementally throughout the project.

The initial list of media sources was composed by a manual review of (1) Google, to find the most-read sources by country; (2) Wikipedia, to check for controversies in reporting; (3) Media Bias/Fact Check, to check for bias in reporting. Next, the researchers involved in the project polled the different research and operational partners of the project per region to review and expand the list of media outlets. The list was also amended and expanded by local project partners in each country. Currently (late 2023), each country has about 10-20 approved media sources, although some countries are currently somewhat under-represented in the system (e.g., Cyprus has 6 sources, Lithuania 8 and Malta 6), which is due mainly to the smaller size of these countries.

The media sources are accessed by the monitor through a commercial News API that offers keyword search functionality across sources. For keywords, we use a list of about 40 expressions related to firearms incidents (e.g., shooting, shots fired) that are composed manually by the firearms researchers involved in the research project. Expert knowledge and in depth understanding of a specific topic – firearms violence – proves to be crucial in developing the list of keywords.

The **crawler** uses the News API to discover recent articles that are probably related to firearms incidents. Every 15 minutes, it searches for articles that contain one or more of more than 40 relevant keywords (e.g., 3D-printed gun, firearms seized, shot to death) in one or more than 20 official EU languages. It continually looks for new articles published during the past seven days. Both the list of relevant keywords (and their translation) and the media sources that are searched were determined manually by the research consortium. Throughout the development of the monitor, **the research team added or deleted keywords from this list, in this way building on the continuous follow-up of the articles that were identified**. For example, the keyword ‘explosion’ was removed from the list because it resulted in a substantial number of articles on gas explosions in the monitor, which did not fall within the scope of the monitor.

In addition, only local articles are observed. For example, for articles written in Dutch, only those published by manually approved local sources (e.g., vrt.be) and which discuss events in the relevant local region (e.g., Brussels) are observed. Articles in Belgian media outlets reporting on incidents in France or Spain, for example, are therefore not included in the monitor. For each article, an English summary is generated on the fly using ChatGPT; this is useful to both human reviewers and AI prediction systems for English, which are usually several steps ahead of the technology available for languages with fewer speakers.

1.1.2 Developing the classifier: 'junk' or 'no junk'?

Since language is ambiguous and words derive their meaning from context (e.g., shot his wife ↔ shot at goal; grenade attack ↔ knife attack), the database is likely to contain false-positives or 'junk' articles. The project has shown that about one-third of all incoming articles are irrelevant. In order to filter out these articles, a **classifier** was developed during the project which predicts whether an article originally identified is relevant or not. This process was also guided by the project's working definition of what constitutes 'firearms violence'. As a consequence, the following cases were not considered as being included in this definition:

- articles referring to explosions and explosives, nor those referring to fireworks and/or pyrotechnics;
- no shooting of animals;
- (legitimate) shootings by police forces;
- (murder-)suicide in which the suicide was committed with a gun, whereas a murder occurred with a weapon other than a firearm (e.g., a knife).

Consequently, such media articles are, for our purposes, labelled as 'junk' and are therefore not shown in the incident monitor.

The first list of keywords that were used to identify relevant articles was developed using the data collected by the Flemish Peace Institute in the context of Project TARGET and by the University of Leiden.[a] Both datasets consisted of media articles on gun violence incidents that occurred in Flanders and the Netherlands respectively. The researchers collected these media articles manually and therefore they were all effectively covering incidents of firearms violence. In total, a combination of these datasets resulted in a set of 4,500 Dutch gun-violence articles. A **word frequency analysis** was then conducted on these articles. Expressions that often occur in these articles and which do not occur often in

[a] In the context of project TARGET a dataset was gathered manually in which all cases of lethal gun violence reported in the Dutch-speaking media in Flanders and Brussels during the period 2010–2020 and all cases of non-lethal and lethal gun violence during 2018–2020 were included. A total of 2,606 media articles that covered 1,167 incidents of gun violence were included in this database (Duquet, N. & Vanden Auweele, D. (2021), *Targeting gun violence and trafficking in Europe*. Brussels: Flemish Peace Institute).

articles about other topics (e.g., gunned down) were retained. This exercise resulted in a list of about 1,500 words and expressions, which are included in the so-called **ontology**. This list was subsequently translated by experienced native speakers of EU languages via the University of Antwerp's translation service, Linguapolis. This list is stored as a Google Sheet in the cloud. This has the advantage that multiple annotators can collaborate on fine-tuning and expanding the same resource.

This process of identifying relevant key words has been repeated iteratively throughout the project as the monitor analysed more articles; this was done with a view to finding relevant expressions that are not yet part of the ontology. The research team also added about 1,000 expressions manually, mostly dealing with the various ways in which ammunition calibers are reported on and also references to age and gender. We also queried WikiData to collect the names of firearms and their ammunition. Each expression in the ontology was used as a keyword in Google News to check how many articles related to firearms it yields and whether a synonym of an expression might yield more. These later entries into the ontology were translated by Google Translate, a manual review of the most frequently occurring expressions also being performed.

The ontology currently contains about 2,500 entries. The research team subsequently classified and labelled the various words that were included in the ontology, as shown in Figure 2. Early on, **this ontology functioned as our first classifier (dubbed 'M1'),** which the monitor uses to predict whether other incoming articles report on firearms incidents or not, on a scale ranging from 0 to 100% (or ☆☆☆☆☆ to ★★★★★). The prediction based on this ontology takes into account the total number of labels that appear in an article: the greater the number of labels that are identified, the higher the score.

Figure 2: Ontology of firearms incidents

Slot	Label	Example	#
CONTEXT	CRIME	*crime, criminal, crook, gun violence, gun running*	105
	CRIME.GANGS	*drive-by shooting, gangs, gang member, gang violence, turf war*	40
	CRIME.DRUGS	*drugs, drug addict, drug dealer, pot form, under the influence of **	74
	CRIME.THEFT	*steal, steals, stolen, swindled, thief*	12
	CRIME.ROBBERY	*armed robbery, mugging, rip deal, rob, robbery homicide*	36
	CRIME.BURGLARY	*burglar, burglary, home invasion, home invader, trespassing*	14

	CRIME.HOSTAGE	*held hostage, kidnapper, kidnapping, tied up, tiger kidnapping*	21
	CRIME.MURDER	*killer, killing spree, man kills *, murderer, murder attempt*	53
	CRIME.TERRORISM	*bomb threat, explosive belt, school shooter, school shooting*	34
	CRIME.THREAT	*brandishing a gun, threat, threaten, threatened at gunpoint*	19
	CRIME.TRADE	*arms cache, arms trafficking, illicit firearms, serial number removed*	22
CONTEXT	POLICE	*law-enforcement officers, police officers, police*	81
	POLICE.CASE	*murder case, murder suspect, person of interest, suspected*	57
	POLICE.CALL	*manhunt, police intervention, police pursuit, returned fire*	46
	POLICE.RAID	*firearms were seized, gun was found, seized, seizure*	22
	POLICE.ARREST	*custody, detained, indictment, incarcerated, was arrested*	63
CONTEXT	DOMESTIC	*romantic relationship, shot himself, shot his wife, son-in-law*	68
	DISPUTE	*heated argument, hooligan, neighbour dispute, traffic dispute*	27
	DISPUTE.LOVE	*adultery, cheating love, love rival, love triangle*	4
CONTEXT	SPORTS	*Champions League, first-person shooter, gun collection, hunter*	67
CATEGORY	EVENT	*rip deal, robbery, robbery homicide, search and seizure*	192
	EVENT.MOVE	*cocaine shipment, drug shipment, flees, pursuit*	8
	EVENT.SHOOT	*grenade attack, shoot-out, shooting, shots fired at house*	62

CATEGORY	ACTION	*fist punch, invaded, killed, kills, pointed a gun, taken hostage*	122
	ACTION.MOVE	*escaped, fled, ran for her life, ran off, took to the streets*	37
	ACTION.SHOOT	*fired, fired at random, fired dozens of, opened fire, riddled*	97

CATEGORY	PLACE	*Dubai, India, Kosovo, Ukraine*	60
	PLACE.HOME	*apartment, family home, porch, quiet residential area*	15
	PLACE.WORK	*bank office, petrol station, gas station, night shop, pharmacy, shopping mall*	25
	PLACE.SCHOOL	*elementary school, high school, university*	4
	PLACE.STREET	*alley, carpool, ditch, highway, parking lot, pavement sidewalk*	11
	PLACE.PUBLIC	*bar, eatery, hotel, nightclub, restaurant, metro station*	20

CATEGORY	TIME	** years ago, last month, last year, over the weekend*	16
	TIME.AM	*around * am, early morning, Saturday morning*	13
	TIME.PM	*around * pm, late afternoon, Saturday evening*	28

CATEGORY	EFFECT	*arrested, at large, caught, detained, handcuffed*	54
	EFFECT.INJURY	○○○ *unharmed, uninjured*	11
	EFFECT.INJURY	●○○ *transferred to the hospital, was injured, were injured, wounded*	31
	EFFECT.INJURY	●●○ *life-threatening injuries, seriously injured, mortal danger*	35
	EFFECT.INJURY	●●● *is dead, died, was killed, were killed, shot to death*	41

CATEGORY	PERSON	** people, *-year-old, adult, boy, girl, gun nut, husband, wife*	216
	PERSON.PERPETRATORS	*drug dealer, gangster, gunman (*), kidnapper, robber, suspect*	139
	PERSON.VICTIM	*man shot down, several casualties, the victim*	48
CATEGORY	OBJECT	*hoodie, jogging pants, mask, sneakers, weed, XTC*	65
	OBJECT.VEHICLE	*cash transport, getaway vehicle, motorcycle, trunk*	28
CATEGORY	PROPERTY	*3D-printed, 9-millimetre, bizarre, cold-blooded, tragic*	46
	PROPERTY.NUMBER	*one, two, three, …*	10
	PROPERTY.COLOR	*Black, white*	2
CATEGORY	WEAPON	*crossbow, knife, Canik, Glock, CZ*	
	WEAPON HW	***heavy weapon**, anti-aircraft gun, gun turret*	
	WEAPON LW	***light weapon***	
	WEAPON LW.SA	***firearm**, gun, guns, ghost gun, small arms*	
	WEAPON LW.SA.HG	***handgun**, hand guns, Colt 45*	
	WEAPON LW.SA.HG.R	***revolver**, Magnum, Webley Mk !*	
	WEAPON LW.SA.HG.P	***pistol**, * mm gun, Glock 17, Glock 19, Sig Sauer P320*	
	WEAPON LW.SA.LG	***long gun**. hunting rifle*	
	WEAPON LW.SA.LG.C	***carbine**, FGC-9, Hi-Point 4095, Uzi*	
	WEAPON LW.SA.LG.R	***rifle**, long rifle, AR-15, Steyr M1888/05, Winchester rifle*	
	WEAPON LW.SA.LG.S	***shotgun**, riot gun, SPAS-12, Ithaca Auto & Burglar*	
	WEAPON LW.SA.LG.M	***machine gun**, AK-47, AK47, AK 47, Kalashnikov*	

WEAPON LW.SA.LG.U	***submachine gun***, *machine pistol, HK UMP 9, MP38*
WEAPON LW.SA.NL.S	***starting gun***, *alarm gun, alarm pistol*
WEAPON LW.SA.NL.A	***airsoft***, *BB gun, gas pistol, pellet rifle*
WEAPON LW.SA.NL.R	***replica***, *fake gun, fake pistol, toy rifle*
WEAPON LW.EX	***explosives***, *grenade, Molotov cocktail, rocket launcher*
WEAPON LW.FW	***fireworks***, *pyrotechnics*

All in all, the system can identify and label 2,500+ common multilingual word combinations related to gun violence, along with 1,000+ firearms denominations (e.g., SPAS-12), although the latter rarely occur in news articles.

In a second phase, an ML **algorithm** was implemented to take over the role of classifier and to predict the relevance of the articles identified in the selected media sources. The monitor uses the ML algorithm stochastic gradient descent (SGD) with character and word *n*-gram features from the first paragraphs of text. This algorithm has been proven to work well in multilingual cases, having trained on 50,000 articles observed by the monitor at that time. We also used isotonic regression to improve the predicted star ratings.

In addition, we improved the algorithm further by having the research team conduct manual controls. Throughout the duration of the project, human moderators regularly and systematically identified irrelevant articles and edited incorrect metadata. For a period of two months, all the research partners were involved in reviewing new articles on gun violence in all the EU Member States and in assessing the relevance of articles that were included in the monitor. During this process, **about 15,000 articles detected by the monitor were controlled manually and flagged as irrelevant.**

For example, the analysis showed a number of false positives in the database relating to gas explosions. This is because we explicitly searched for articles containing the keyword ‘explosion’ with the intention of capturing grenade attacks. However, gas explosions were then also picked up, and as they are written in a similar style as reports on firearms (e.g., mentioning victims, injuries, neighbours hearing loud bangs, police response), these were sometimes picked up by the classifier. In the example, we replaced the search keyword ‘explosion’ with ‘grenade attack’, we expanded the ontology with entries such as ‘gas explosion’ and marked these as indicators of false positives. We also retrained the classifier (first M1, then M2, finally M3) with new examples (articles about gas explosions manually deleted by moderators), which led to more reliable predictions. During the analysis, we dealt with many of these fine-grained issues. Note that, in general, in about 9/10 cases it is easily apparent whether an article belongs in the database to human beings as the classifier or not. The remaining 1/10 are articles that are more unclear and which often lead to disagreement even between human beings (e.g., a breaking news report where it is not yet known whether a gas explosion accident or a drug-related grenade attack occurred).

Because, generally speaking, the manually reviewed control set of 15,000 articles (mentioned earlier) is too small to train an ML prediction model on, the research team decided to integrate ChatGPT into the classifier. **The current classifier (dubbed M3) in other words combines the algorithm and ChatGPT to evaluate the relevance of incoming articles identified** in the first step. ChatGPT is instructed to check for the overarching topics in order to identify relevant firearms violence in the media sources that were listed:

> 'firearms incident', 'firearms trafficking', 'firearms confiscated', 'firearms discovery', 'firearms stolen', 'firearms seized', 'munition seized', 'illicit firearm', 'illegal firearms', 'shooting', 'shootout', 'shooter', 'shots fired', 'shot dead', 'shot themself', 'gun violence', 'gunshots', 'gunshot fired', 'gunshot wound', 'gunshot death', 'warning shots', 'arrested at gunpoint', 'threatened at gunpoint', threatened with a gun', 'carrying a loaded gun', 'grenade attack', 'fake pistol', 'fake gun', 'toy gun', 'air gun', 'imitation gun', 'homemade gun’.

The reliability and validity of this most recent classifier were tested by asking ChatGPT to review another 150,000 articles. As it turned out, by comparing the 15,000 human reviews (see earlier) to ChatGPT’s predictions, the statistical agreement between human beings and ChatGPT is above 80%. Human beings and ChatGPT agree on clear-cut cases (i.e., articles that ChatGPT assigned a probability of < = 0.1 or > = 0.9) to approximately 95% of the time. In cases that are not clear cut, ChatGPT often does better than human beings – for example, by discarding articles about wildlife control. Afterwards, in order to improve further the reliability of the training data, we manually reviewed and corrected 5,000 cases where human beings and ChatGPT disagreed strongly. This makes ChatGPT’s assessments useful as training data. We proceeded to train a model on the 150,000 predicted labels, which constitutes the current M3 model that is deployed live.

Based on these analyses, it appears that the current classifier is slightly better at filtering out unrelated articles (approx. 93% of them are removed) than it is at correctly identifying firearms incidents (approx. 87% of them are detected). Because a false-negative (therefore the wrongful assessment of an article as ‘junk’) is considered to be less problematic than a false-positive (the inclusion of a non-relevant article in the monitor), articles with a star rating of less than three are currently not included in the monitor. These articles add very little insight, especially because they have been found to have a high human removal rate.

To summarise, the current database shown by the monitor contains about 60,000 articles ranging from January 2019 to December 2023 that relate to firearms incidents in the EU with a predicted rating of ★★★☆☆ or more. Several steps were taken to mitigate false-positives and false-negatives, but continued human oversight is required.

1.2 Clustering different articles on the same incident

A second important step in the monitor is to cluster together media articles that have appeared in different media sources or on different days and which deal with the same gun violence incident. Such clustering of articles is crucial, with a gun violence incident being a unit of analysis rather than a media article. Especially when spectacular incidents take place that have a lethal outcome or are politically motivated, various media sources will report on them. In such cases, media reports will also be published on several days, often adding additional, updated information on the context, the outcome, the persons involved and the firearms used.

The clustering algorithm therefore attempts to group related articles into events automatically. This algorithm uses a **Deep Learning (DL) approach called 'text embedding'**. Text embedding is a sequence of numbers that represent a text document. The *cosine distance* between two number sequences can then be calculated as a value between 0.0 and 1.0 that represents the degree of similarity between these texts. We use the current state-of-the-art model, distiluse-base-multilingual-cased-v2, to create text embeddings for all observed articles.

The current clustering algorithm however has a weakness comparable to 'Whisper Game'. In short, an article from two days ago may be similar to an article from yesterday, and an article from yesterday may be similar to an article from today. But the article from today and the article from two days ago may have nothing in common. This chaining can lead to large groups stretched over time that, as a whole, have nothing in common. This effect can be limited by setting a high threshold score, which essentially means that articles must be more *grammatically* similar than *semantically* similar in order to form a group. This is a constraint of the system. Elaborate tests that were conducted by the research team and the monitor developers indicated that the average similarity is 0.75; this is a baseline measure when human annotators decide to group articles into events by hand. The clustering algorithm also creates groups with an average similarity > 0.75 so that it performs comparably to human beings. This threshold can be changed. For example, lowering it will generate larger, looser clusters with more linguistic variation among the articles (different wording, different information). However, this also increases the likelihood that unrelated articles will become grouped together. Because of the relatively high threshold that was decided to being used, very few cases were identified in which the clustering algorithm incorrectly clustered articles that discussed different incidents. When reports on the same incident were not clustered together, false-negative cases occurred more often, although in still rather limited instances.

New article groups in the monitor are also manually reviewed continuously. Moderators have tools at their disposal to regroup and annotate groups of articles. Because of the high threshold that was used to cluster articles together effectively, the research team clustered several articles manually. Although this manual process helps with developing the algorithm further, it shows equally that continuing manual controls remain necessary.

The clustering algorithm can also reveal links between events over a longer time span, but throughout the development of the monitor and the manual controls that were conducted on the clustering, it appeared that the clustering algorithm works well for articles published in the same time span but not over a longer period. This is, for example, the case when articles are published in the direct aftermath of an incident itself and new articles are published on the legal procedures that take place months or even years after the effective incident.

1.3 Extracting relevant information on gun violence incidents from media articles

The third aspect of the real-time monitor is to extract and present relevant contextual information on a gun violence incident in an automated manner. Media articles can be particularly relevant because they often include information on the context of an incident, the outcome (lethal outcome, injuries, threats) and the characteristics of the perpetrator(s) and victim(s), such as their age and gender. More specifically, the incident monitor tries to find information on the following elements: type of incident, motive, age and gender of perpetrators and victims, their relationships, the number of persons killed, injured and arrested, whether a suicide, whether shots were fired, whether an explosion occurred, whether on trial, whether weapons were seized, and the firearm type and the ammunition used. This information is integrated and visualised in the monitor at the incident level.

Our first approach was to work with the broad list of keywords to identify certain relevant contextual elements in the media articles. The **ontology of keywords**, described in 1.1.2, was therefore also used to identify relevant contextual keywords in the selected articles. This system is called the **tagger**. The tagger assigns labels to words and word combinations in an article. For example, shooter will be labelled as a PERSON.PERPETRATOR. Each word or word combination can have multiple labels. More specifically, the tagger will assign a category (e.g., shooter = PERSON, shooting = EVENT, shoot = ACTION, shot = EFFECT), a context (e.g., drive-by shooting = CRIME.GANGS) and various other fine-grained details that signal a person's age, gender or motive, the severity of injuries, the type of firearm(s), components, ammunition calibre, and so on. This AI task is also known as 'entity extraction'. The general idea is that labelled entities (persons, objects, places, events) can then be aggregated into structured and quantifiable insights ('slot-filling').

Well-known natural language processing (NLP) techniques such as dependency parsing have been developed to resolve this problem: a parser automatically identifies word types and their relationships in a given sentence. However, whereas many robust parsers have been built for English, there is a historical lack of tools for under-represented languages, and we need to analyse news articles in all of the EU's 22 languages. An additional challenge is co-reference resolution, an NLP approach that is used to identify those expressions that refer to the same entity, as in: 'The man was arrested. He is 30 years old' (*man → he → 30*). The available NLP tools for co-reference resolution are typically also less robust due to the ambiguous nature of the task (e.g., 'The policeman shot the suspect. He was detained.').

One of the more challenging tasks is to identify reliably the **age & gender** of perpetrators and victims mentioned in news articles. Some mentions such as 'the male suspect (30)' are straightforward, but many other cases are more complex. For example, in 'the man who shot his wife is 30 years old', the perpetrator (*man who shot*) and the victim (*his wife*) have to be inferred from sentence structure based on the relational verb *shot*, and the age (*is 30*) needs to be attributed to the perpetrator instead of the victim (*his wife is 30*).

Because of the inherent challenges in using the ontology to extract contextual information from an article, especially with there being 22 official languages in the EU, a second method was implemented. In this second method, LLMs are used to process the information included in the identified media articles. With the recent upsurge in such efficient, fast and cost-effective LLMs, part of the manual research can be augmented by **virtual assistants** (chatbots). Once again, we used OpenAI's ChatGPT.

Even though LLMs are sometimes prone to factual errors ('hallucinations') and Europol has warned about a 'grim outlook' for their misuse in cybercrime and harmful propaganda,[8] ChatGPT also excels at several tasks that were previously challenging in NLP, including machine translation (MT), text summarisation and named entity recognition (NER). ChatGPT is integrated into our monitor as a real-time virtual assistant that identifies types of crime, place, people, weapon and ammunition, plus their relationships and context, in all the EU languages. The early-stage adoption of new LLMs has greatly advanced the reliability and usefulness of our system. To name only a few improvements:

- we are better able to differentiate between accidents, crimes and court cases;
- the detection of illicit firearms seizures has improved;
- the coverage of age and gender detection for perpetrators and victims increased from 35% to > 80%; and
- GPS localisation has become more granular.

Manual tests conducted by the research team indicated that ChatGPT is able to extract much more relevant and correct information from the media articles. More specifically, an approach was developed in which the Chatbot was, for example, prompted with: 'Read the article, name the perpetrator, give the perpetrator's age as a number.' This also works for NLP-scarce languages such as Bulgarian. ChatGPT will answer many additional questions, such as the **type** of crime (e.g., burglary or robbery), the **motive** of the perpetrator(s), the **relationship** (if any) between the perpetrator and the victim, the number of victims, the weapons used, the number of weapons **seized**, the number of shots fired, and whether the article is breaking news or covering the court case of an event that occurred in the past. In many cases, ChatGPT can also provide the GPS coordinates of the event down to the street level.

Our main challenge was to get ChatGPT to answer in exactly the way in which we wanted it to be included and reported in the monitor. This is also called 'prompt engineering'. By default, ChatGPT responds in a conversational tone by writing fluid sentences, but we benefit more from easily quantifiable data such as lists of key–value pairs, that is, `AGE = 30` instead of 'according to the article the perpetrator appears to be thirty years old', which would again require dependency parsing. In our case, we instructed ChatGPT to respond in JSON, an open standard data format for storing key-value pairs. We also instructed ChatGPT to respond only with predefined labels. For example, the type of crime label could be ACCIDENT, CRIME.BURGLARY, CRIME.ROBBERY, CRIME.MURDER, and so on, but nothing else. This task requires some attention, but once the prompt and any post-processing are stable, the task can be applied infinitely to every incoming article.

For this reason, both methods – the ontology and ChatGPT – currently exist next to each other, although a hierarchy has been implemented. The basic annotation continues to be done via the ontology, as it works much faster. This initial annotation is then checked against the information derived from the ChatGPT method. Throughout the continuous manual controls the method has been found to be highly reliable and accurate in extracting useful information from the articles – provided that that information is available. For some specific technical specifications such as a firearm classification the ontology still appears to identify such information in a more reliable and valid manner in comparison to ChatGPT. **Using both methods in the current monitor in other words increases the comprehensiveness and correctness of the information extracted from the media articles**.

Developing an automated report generator

The monitor also enables automated monthly trend reports to be generated; these reports are based on the information in the articles that were included in the monitor.

The trend reports are derived from the aggregation of fact sheets and they provide key insights into firearms violence over time – for example, the number of incidents per month or per year, hot spots, the main firearms used, the correlation between type of crime and firearms used, the correlation between type of crime and perpetrator's age, and the correlation between a victim's gender and the perpetrator's motive.

These statistics are essentially sums generated in a given time frame. For example, for 50 groups of articles in the month of May, if 25 groups are labelled CRIME.ROBBERY, the report will say that 50% of the incidents observed were robberies. Among the 50 groups, it is then also possible to find, for example, the minimum and maximum ages of the perpetrators. The system pays more attention to sums that can be verified by 'consensus' (averages) and 'corroboration' (more articles with the same labels).

These statistics are integrated and published in a template that was developed by the research team with a view to presenting the most relevant data in a structured and systematic manner.

2 Concluding remarks, lessons learned

This second and final section of this methodological report contains some concluding remarks and some lessons learned from the development of an online monitor of firearms violence incidents in the EU, using a variety of AI techniques.

The different AI techniques used in the development of this monitor have proved to be an added value in supporting the building of a better understanding of firearms violence across all EU Member States. In this regard, this monitor can contribute to a better intelligence picture of the different types of firearms violence committed in the EU and the contextual characteristics of this phenomenon. These insights could be relevant to both policymaking and operational law-enforcement activities.

Besides the **inherent limitations** that have been exposed in using media articles as the main source of information on firearms violence in the EU (which were discussed in the previous section), the development and implementation of the monitor has shown that further limitations still exist. **First**, language remains a key limitation, especially given our focus on the EU with its 22 languages**.** The reviewers in the research team are not native speakers of languages such as Romanian or Bulgarian and therefore they were not able to check the validity of (predictions for or generated summaries of) articles written in these languages to the same extent as other languages such as Dutch, English, French and German. **Second**, it can be concluded that the current AI systems are still not completely able to classify the identified articles correctly as either 'junk' or 'no junk' with 100% reliability. However, it is important to mention that human beings also make mistakes. And, **third**, the clustering of different articles covering the same incident remains an equally important challenge. As the 'context window' of LLMs (i.e., their short-term memory) continues to grow, so will we be able to ask them to examine more articles in bulk and group them meaningfully. These limitations should be taken into account when using the information in the monitor.

Despite these limitations, though, the combination of unique and substantial expert knowledge of both firearms violence, the use of AI techniques to identify and analyse online data in the project consortium, the substantial time invested in developing, testing and adapting the AI methods, and the implementation of newly developed AI techniques has resulted in a unique technology as this point in time. In particular, this project has managed to implement LLMs in the same month as they become commercially and computationally viable. **In combination with the unique, uncommon human expertise on firearms, this technology and the monitor that has resulted from it seem to be unique at this point.**

The development of this unique technology has also made it clear that the role of human intervention cannot be overstated. First, sufficient and in-depth **expert knowledge** on firearms violence throughout the full development process, but in particular at the monitor's inception, has proved indispensable. Especially with regard to the original development and selection of key words and the identification of relevant articles, such expert knowledge has appeared to be crucial. Building on the often-heard claim about the added value of AI – '**garbage in, garbage out**' – this project has shown that such expert knowledge is crucial to increasing the quality and relevance of the articles that would be selected by the algorithms. Second – and this is runs as a red thread throughout this report – constant human oversight and control remain crucial and necessary. Continued training of the algorithms in combination with continuous technical controls by humans to assess the reliability and validity of the outcome of the subsequent steps – identification, clustering and data extraction – is expected to remain necessary. More generally, human oversight will become increasingly important as AI becomes more closely integrated into high-stakes decision-making and policy recommendations, as is also evidenced in the recent EU AI Act.[9]
Automating the data collection on firearms, therefore, has important advantages; but at the same time human control is considered to be necessary. However, for our purposes, this human element is expected to decrease over time as the various AI methods are improved during the collection of new data. However, it has become apparent during the development of the monitor that basic technical opportunities for human intervention remain both necessary and desired.

Besides the continuing improvements being made to the current scope and focus of the monitor, the technology that has been developed and the expertise to develop and implement this technology could possibly also be used and applied in other contexts and on other types of data. The technology could, for example, be used to expand the geographical scope to other relevant areas such as the regions bordering on the EU (e.g., Ukraine, the Western Balkans, the United Kingdom), Latin America, North America, and so on. Moreover, it could possibly also be applied to types of data other than media articles. Using the technology and the AI techniques that have been developed on 'raw' police data (such as case records) or customs data would, for example, be an interesting avenue to explore.

However – and this is important to reiterate – it cannot be assumed that the algorithms which were developed during this process can be transferred to other geographical regions or to other types of data as is, without adaptation. On the contrary, doing so will necessitate a substantial involvement of expert human knowledge, investment in expertise and making the time to develop and implement a meaningful, reliable and valid instrument. The development of this online firearms incident monitor has shown time and again that such human involvement and control are crucial and indispensable conditions, as has been stressed on numerous occasions in this report. **Although various AI techniques clearly offer**

substantial advantages in gathering and analysing information – that is, information on incidents of firearms violence – in a very efficient and effective manner, the role of the human factor in contributing to the success of this monitor cannot be overstated. Continuous human engagement in and control of the different AI processes that drive this monitor will therefore remain a necessary condition of guaranteeing its relevance.

Endnotes

1 Communication from the commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions (2020), *2020–2025 EU Action plan on firearms trafficking*, COM(2020) 608, 24/7/2020, Brussels, https://eur-lex.europa.eu/resource.html?uri=cellar:65f0454e-cfef-11ea-adf7-01aa75ed71a1.0003.02/DOC_1&format=PDF.

2 Duquet, N. & Vanden Auweele, D. (2021), *Targeting gun violence and trafficking in Europe*, Brussels: Flemish Peace Institute.

3 Jingfors, S., Lazzano, G. & McConnell, K. (2015), A media analysis of gang-related homicides in British Columbia from 2003 to 2013, *Journal of Gang Research*, 22:2, pp. 1–17.

4 Krüsselmann, K., Rabolini, A. & Liem, M. (2021), *The illegal use of hand grenades in the Netherlands: 2008–2021*, The Hague: Leiden University.

5 Liem, M.C.A. & Koenraadt, F. (2007), Homicide-suicide in the Netherlands: A study of newspaper reports, 1992–2005, *The Journal of Forensic Psychiatry & Psychology*, 18:4, pp. 482–493.

6 South Eastern and Eastern Europe Clearinghouse for the Control of Small Arms and Light Weapons (2022), *Armed violence monitoring platform*: https://www.seesac.org/About-the-platform/.

7 De Labbey, Q., Vanden Auweele, D. & Duquet, N. (2022), *Firearm trafficking and gun violence in Belgium*, Brussels: Flemish Peace Institute, p. 7.

8 ChatGPT – the impact of Large Language Models on Law Enforcement | Europol (europa.eu).

9 Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), European Parliament, P9_TA(2023)0236, 14 June 2023, TA (europa.eu).