Forensic Science International: Digital Investigation 31 (2019) 200889

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

Worldwide analysis of crimes by the traces of their online media coverage: The case of jewellery store robberies

Giulia Margagliotti*, Timothy Bollé, Quentin Rossy

Ecole des Sciences Criminelles, University of Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 2 July 2019
Received in revised form
2 October 2019
Accepted 30 October 2019
Available online 4 December 2019

Keywords:
Jewelry store robbery
Online news
Pattern detection
Spatiotemporal analysis
Crime analysis
Monitoring methodology

ABSTRACT

This empirical study aims to determine whether online media coverage can be used to gather intelligence on specific crimes worldwide. The quality of online news is evaluated as an indicator of the worldwide distribution of jewelry store robberies. This phenomenon was selected because evaluating the risk of criminal events at the global level is a challenge for private companies, who need to settle and prioritize protection strategies to determine the actual risk within each country. Online media coverage is thus scrutinized for its ability to reveal spatiotemporal trends of this phenomenon. Based upon a dataset of online news gathered between 2015 and 2017 from the news aggregator website EMM (Europa Media Monitor – NewsBrief), the results show that online news may be a cost-effective method to analyze risks worldwide — though a cross-check with different data sources is still necessary to validate its accuracy. The developed approach shows that (1) while a multilingual approach is required, (2) cases can be detected and automatically classified with good accuracy; (3) moreover, dates and countries of published news articles are generally reliable indicators of the actual times and places of the events, which reduce the need for complex text analysis methods. This study demonstrates how a simple monitoring approach can be used to support the worldwide spatiotemporal analysis of serious crimes such as jewelry store robberies.

Introduction

In today's society, the media have become an essential component of everyday life. The media are the most powerful tools of communication, and with the advent of the Internet, they are constantly and asynchronously available from any connected device. Online news holds different functions — from aggregating stories and information related to specific topics to increasing public awareness and shifting attitudes on different societal issues. News is not a direct source of criminal activities. Nevertheless, it traces the media coverage of criminal activities. As such, news is scrutinized for its potential as an indirect indicator to monitor crime trends. The question addressed in this contribution is whether online news may be used as a valuable source.

Since they are broadcast in online spaces, digital forensic processes may be used to collect, collate and analyze online news. This paper aims to present a methodology based on machine learning techniques to automate the processes and evaluate the potential of metadata analysis as a valid proxy of the actual date and time of the occurrence of the event. Indeed, analyzing the spatiotemporal distribution of crimes around the world based on media coverage involves (1) accurately detecting and filtering relevant news and (2) extracting valid indicators of their time and location of occurrence. If the content of the news is analyzed to extract spatiotemporal data, can the metadata of the trace itself be used (i.e., the date and location of the publication)?

To evaluate the potential of this approach, the phenomenon of jewelry store robberies is used as a promising example. Indeed, it is a major and worldwide problem for the luxury sector because of its financial and human impacts. Taking adequate security measures to cope with the actual risks and harms at the local level is not a straightforward approach. Global trends of the phenomenon that lead to global security standards may be defined to ensure minimum security requirements. Nevertheless, the actual risks may greatly vary according to the situation of the national/local area. The question is therefore how international firms can set up adequate security measures and cost allocations at the local level. To reach this objective, an intelligence-led approach may be used. The basic method of crime intelligence has become a classic

* Corresponding author.
E-mail address: giulia.margagliotti@unil.ch (G. Margagliotti).

https://doi.org/10.1016/j.fsidi.2019.200889
1742-2876/

approach in policing culture and is broadly documented (Atkin, 2000; Ratcliffe, 2016). This method aims to timely turn raw data into appropriate intelligence that supports decisions; but how can such an approach be settled at an international level?

The implementation of the approach requires appropriate data, and several sources are considered. The first source is companies' databases, which contain information about their own victimization. Creating a joint database may be difficult since it requires collating data from separate firms or stores and distinct security managers and countries. Otherwise, the risk assessment is based on the incidents experienced by a single victim, which leads to a limited account of the global problem. Secondly, official data from law enforcement can also be used. However, the main issues in this case are the accessibility and the gathering of the data at the international level. Thus, open source data, and more specifically the analysis of online news, are scrutinized as an alternative.

The use of online news to analyze criminality

In the news, crime receives ample coverage compared to other types of events. Serious crimes are often emphasized and generally considered more newsworthy than less serious crimes (Graber, 1980). Graber finds that crime-related news accounts for 22% to 28% of stories in the newspapers analyzed.

Over the last ten years, much criminological research has been conducted using information from the news. Some studies used the news to assess how the media cover crimes. For instance, Taylor (2009) presents the analysis of an American newspaper addressing female victims of murder to evaluate the media interest and to identify patterns of victim blaming in the media. Another aim is to use the media coverage to analyze crimes. For this latter aim, the approach is based on the premise that criminal activities, in particular serious crimes, attract the attention of the media. If the crimes are covered by the media that broadcast the information online, then the news can be used as a source of information to perform quantitative and qualitative research. For instance, the prevalence of cases on a global scale may be evaluated and trends among countries may be compared. For instance, Sharma et al. (2015) proposed a method to identify the most secure path between two places within a city. To do so, they used police reports and online news to estimate the localization of the hotspots in the different areas of New Delhi. The description of the cases may also be used to analyze and compare the characteristics of the crimes, victims and offenders. For example, Aniello and Caneppele (2018) conducted an exploratory study on the resale of stolen goods on the Internet. They focused on cases reported by online news within a two-year time frame and examined the characteristics of the market as well as the most used methods for fencing. Social reactions may also be compared to identify trends and best practices. Consequently, online media coverage may be an interesting indicator to perform a global analysis, where the comparison of official data (i.e., police and judicial datasets) and victimization surveys are known to be difficult to obtain and expensive.

However, the indicator can only be valid if the crimes are actually well covered by the press. Thus, the development of crime analysis methods based on online news should consider the results of the studies on how the media cover crimes.

The work of Ericson et al. (1991) is one of sociology's classic studies on the subject. It illustrates the decision-making process that leads the media to publish certain stories and how they are depicted. News is seen as a form of social control with regard to issues of crime, law and justice, which is evidenced by the dominance of judicial stories in the news. Five components of representation in news communication are identified: visualizing, symbolizing, authorizing, staging and convincing.

Marsh (1991) conducted an analysis of 36 crime articles published in the United States between 1960 and 1988 and 20 studies conducted in several countries during the same time. He found an overrepresentation of violent and interpersonal crimes and an underreporting of property offenses, compared to official statistics. Chibnall (1977) and Jewkes (2015) identified the specificities of the criminal acts reported by the media. They (1) are visible and spectacular acts, (2) have a political or sexual connotation, (3) are connoted by a peculiar pathological aspect, (4) entail a high-linked risk, (5) involve a cultural or spatial proximity, and/or (6) involve children. As a result, not all crimes are monitored by the media but rather those that are violent and able to astound or shock the public. For instance, Coscia and Rios (2012) analyzed drug trafficking, Arulanandam et al. (2014) studied theft cases reported by online newspapers from New Zealand, India and Australia, and Das and Das (2017) implemented a method to analyze crimes against women in India.

In sociology, the class-dominant theory argues that the upper class controls the economy through the corporate community and the media reflect and project the view of a minority elite, which controls the economy (Barlow and Decker, 2010). The media evidently profit mostly from advertisements, and the bulk of these profits are from multinational companies, political parties and wealthy individuals who advertise on media platforms. As such, news outlets may be reluctant to publicize negative stories about them. It can also be argued that local ownership and control of news media are beyond the reach of large corporate offices elsewhere and that the quality of the news largely depends on the journalism. For instance, numerous environmental causes receive full media coverage and support.

Since the majority of news articles are online and free of charge, they are easily accessible, and they are collated through aggregators. The key issue is thus to detect the cases among the mass of online news and extract the information relevant to the analysis. To support the process, previous studies have suggested the use of data mining and machine learning (ML) approaches. Coscia and Rios (2012) developed a method called making order using Google as an Oracle (MOGO) by exploiting online news and blogs considered trustworthy to acquire information on the areas of activity as well as the modus operandi of Mexican criminal organizations. This method helps to gather intelligence about the localization and structure of these groups, which is useful information to set up adequate mitigation strategies. Arulanandam et al. (2014) use ML techniques, such as named entity recognition (NER) and conditional random field (CRF), to identify all words that reference a place in written text. Based on newspapers from Sri Lanka, Jayaweera et al. (2015) present another method. The collected articles are reviewed through a focused crawler and classified with a support-vector machine (SVM) algorithm as belonging to the category “crime” or not. Entities such as names, places, organizations or dates are extracted with natural language processing (NLP) techniques and the duplicates are detected. Dasgupta et al. (2017) proposed an NLP technique that allows the extraction of a maximum of knowledge concerning crime events reported by the media. This information includes, for example, the names of both authors and victims, the nature of the crime, the geographical area, the date and the time of the event as well as the action taken against the criminal. Das and Das (2017) used similarity measures and a clustering algorithm to select the most pertinent characteristics to reveal groups among different Indian regions sharing similar characteristics in crimes against women. Finally, Rohini and Isakki (2016) highlighted the importance of data visualization to detect trends among the collected information. They described the methodology in six steps: web crawling, document classification, entity extraction, duplicate detection,

Table 1
Confusion matrices for the French dataset (left) and English one (right).

French	Predicted			English	Predicted		
Actual		**F**	**T**	**Tot.**	Actual		**F**	**T**	**Tot.**
	F	257	146	403		**F**	319	285	604
	T	46	1649	1695		**T**	37	1730	1767
	Tot.	303	1795	2098		**Tot.**	356	2015	2371

information visualization, crime analysis and prediction.

Furthermore, this type of approach using online news reports has been developed in other fields, and in particular in epidemiology. It is used to monitor the emergence and spread of certain types of infectious diseases in countries where public health authorities have limited resources (Doan et al., 2008; Chan et al., 2010; Mollema et al., 2015).

Most studies focused on the technical developments of the methodology. None of the previous research has focused upon the use of online news to analyze spatiotemporal trends of a criminal phenomenon. To do so, we chose to focus on jewelry store robberies, which were never studied by these means. Our study addresses four questions: Can online open sources be used to gather information on jewelry store robberies worldwide; can risk assessment be made at the local level (i.e., within a country); how does the search language influence the detection process; are the country and the date of the publication of the article reliable indicators of the actual place and time of the event?

It is a first step to evaluate whether the international online media coverage of jewelry store robbery cases allows the production of relevant and timely intelligence that is useful for prevention and risk management.

Methodology

According to the aforementioned studies, a methodology of five stages was established: (1) online news collection, (2) news filtering and case identification, (3) named entity extraction, (4) duplicates detection, (5) spatiotemporal analysis.

Online news collection: French and English keywords were selected for the news article search (*braquage bijouterie, braquage joaillerie, vol bijouterie, vol joaillerie, robbery jewellery store, robbery jewelry store*), and queries were used to gather online news articles published between January 1, 2015, and December 31, 2017, from the news aggregator website EMM. It was assumed that the English language should allow one to cover most countries since the majority of these countries have English media reporting the news. The list of EMM sources showed that the majority of the countries were indeed covered by English media, except for Latin America and some Asian, Eastern European and African countries. Thus, it was not expected to detect cases published by the media of these geographical areas. EMM is a freely accessible platform, gathering about 300,000 articles per day from different international, national and local online news portals from all over the world. The “all these” (terms) type of search was used separately for English and French. All of the keywords entered must appear in the article for it to be collected, although they may appear several times within the text. Two datasets containing the French and English results were generated for further comparisons. Articles were filtered out if they shared the same title and URL and then the source code was retrieved for each news article, allowing the extraction of the full content of the news by targeting specific HTML tags, such as “body” in the first place and “h1,” “h2” and “p” in the second place.

News filtering and case identification: The content of the article previously extracted was processed with a sequence of natural language processing operations. The natural language toolkit (NLTK) Python library (Bird et al., 2009) was used to perform the tokenization, which is the segmentation of the text, the deletion of the stop words and the lemmatization, which is the removal of the inflectional endings from the words and the conversion into their lemma. Then, the processed text was vectorized with term frequency inverse document (TFIDF) vectorizer to calculate the frequencies of the words in the corpus of the document. At this stage, the SVM classifier was trained on a sample dataset, then the supervised learning operation was performed on the whole dataset.

Fig. 1. World distribution of jewelry store robberies between 2015 and 2017 detected in the online news with French keywords (N = 459).

The result was the classification of each news article into two categories: "T," standing for an article covering an actual case of jewelry store robbery generally defined as a theft involving the use of violence, and "F," standing for any other content.

The SVM model was trained with datasets containing 300 articles for each language. The datasets are skewed because they contained many more cases of jewelry store robberies than other contents. In this situation, the F1 score measure was retained to estimate the performance of the model (Jeni et al., 2013), which is on average about 80%. It was calculated with the following formula:

$$F1\ score = 2\frac{PR}{P+R}$$

where P and R represent precision and recall, respectively, and are calculated as follows:

$$Precision = \frac{True\ Positives}{True\ Positive + False\ Positives};$$

$$Recall = \frac{True\ Positives}{nb.\ of\ predicted\ positive}$$

Then, the classification was executed on the rest of the data and manually verified to avoid missing actual cases of misclassified jewelry store robberies. The manual verification of the results also allows using a confusion matrix to evaluate the performance of the automatic classifier for each language.

Named entity extraction: Relevant information concerning the place (city and country) and the date of the event were manually extracted from true cases of jewelry store robbery. The extraction phase was performed in the following manner:

- The short description containing the first lines of the articles was read to find the information of the date and place;
- If the information was missing in the short description, the whole article was read;
- If the information concerning the place was not available, the corresponding variable was left empty;
- If the information concerning the date was not available, the date of the publication of the article was kept;
- If information concerning the time of the event was present, it was added to the corresponding date variable;
- When an article reported more than one jewelry store robbery case, lines were added to the database, and they were completed with information concerning the country and the date of the publication of the article as well as with the actual place and date of the event.

As mentioned above, when the information concerning the date of the event was not available, the date of the publication of the article was kept, which made the calculation of the delta equal to zero. The delta is calculated by counting the number of days between the date of the publication of the article and the actual date

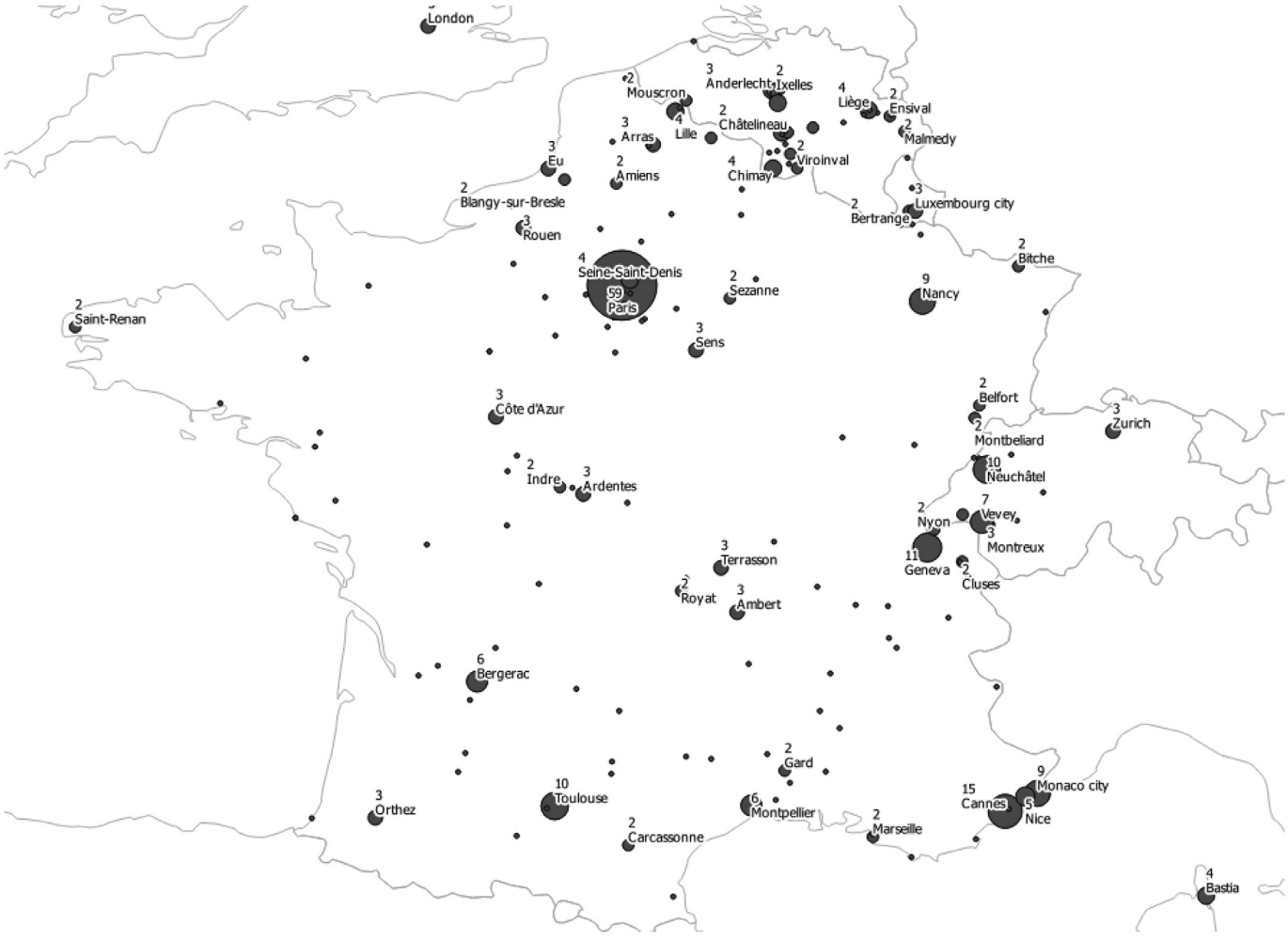

Fig. 2. City-level distribution of jewelry store robberies in France, Belgium and Switzerland (N = 207).

of the event. Nevertheless, this scenario happened very rarely, since the date of the event was almost always mentioned.

Duplicates detection: The articles covering cases happening on the same day in the same city were considered duplicates and counted once. Therefore, on the basis of the information extracted from the previous phase, a unique identifier was finally attributed to each event allowing the detection of duplicates.

Spatiotemporal analysis: At this stage, different visualizations of the data were performed to detect temporal crime patterns and analyze the spatial distribution. Two levels of spatial resolution were used for the analysis: (1) at the country level, where choropleth maps were performed and (2) at the city level, where proportional symbol maps are used. QGIS software was used to create the visualizations. It is worth mentioning that no particular temporal patterns were found, and for this reason, this aspect is not shown or discussed in the results section.

Fig. 3. Country-level distribution of jewelry store robberies, between 2015 and 2017 detected in the online news with English keywords (N = 1058).

Fig. 4. City-level distribution of jewelry store robberies, between 2015 and 2017 detected in the online news with English keywords. (N = 1058).

Results

Search and identification of cases

The news gathering performed with the French keywords collected 4821 articles. After the filtering and extraction of the source code to obtain the articles' text, some articles were filtered out because they were no longer accessible. Thus, N decreased to 2115. The English search gathered 4324 articles. After the aforementioned operations, N decreased to 2375. The fact that this first phase of the methodology amounted to almost as many cases in French may be surprising, as we expected news media publications in English to strongly outnumber French ones. It is thus possible that the French keywords were more precise than the English ones and allowed us to collect more articles related to jewelry store robberies. We can infer the hypothesis that the English keywords were not as appropriate, as certain countries may use different English words to describe the same phenomenon.

At this point, the SVM classification model and the manual check of the results were carried out. The confusion matrices shown in Table 1 were obtained.

In terms of the French dataset, the measure of the F1 score is equal to 0.95. Similarly, for the English one, it is equal to 0.91, which overall shows a good performance of the classifier. This indicates that such technique might be used to automatically filter pertinent news.

Spatiotemporal analysis

Among the 1695 articles identified in the French dataset, 508 did not belong to the time frame under study (i.e., 2015–2017), and 25 articles covered robbery attempts. In the English dataset, among the 1767 articles, 199 were out of the studied time range. These articles were excluded from the analysis.

Furthermore, some articles covered more than a single event, and for this reason, 10 cases were added to the French dataset and 46 were added to the English one.

After the named entities extraction and the detection of duplicates and their exclusion, 459 French articles and 1058 English articles were considered for the spatiotemporal analysis.

Fig. 1 highlights that France and its neighboring countries are the main jewelry store robbery hot spots, according to the data

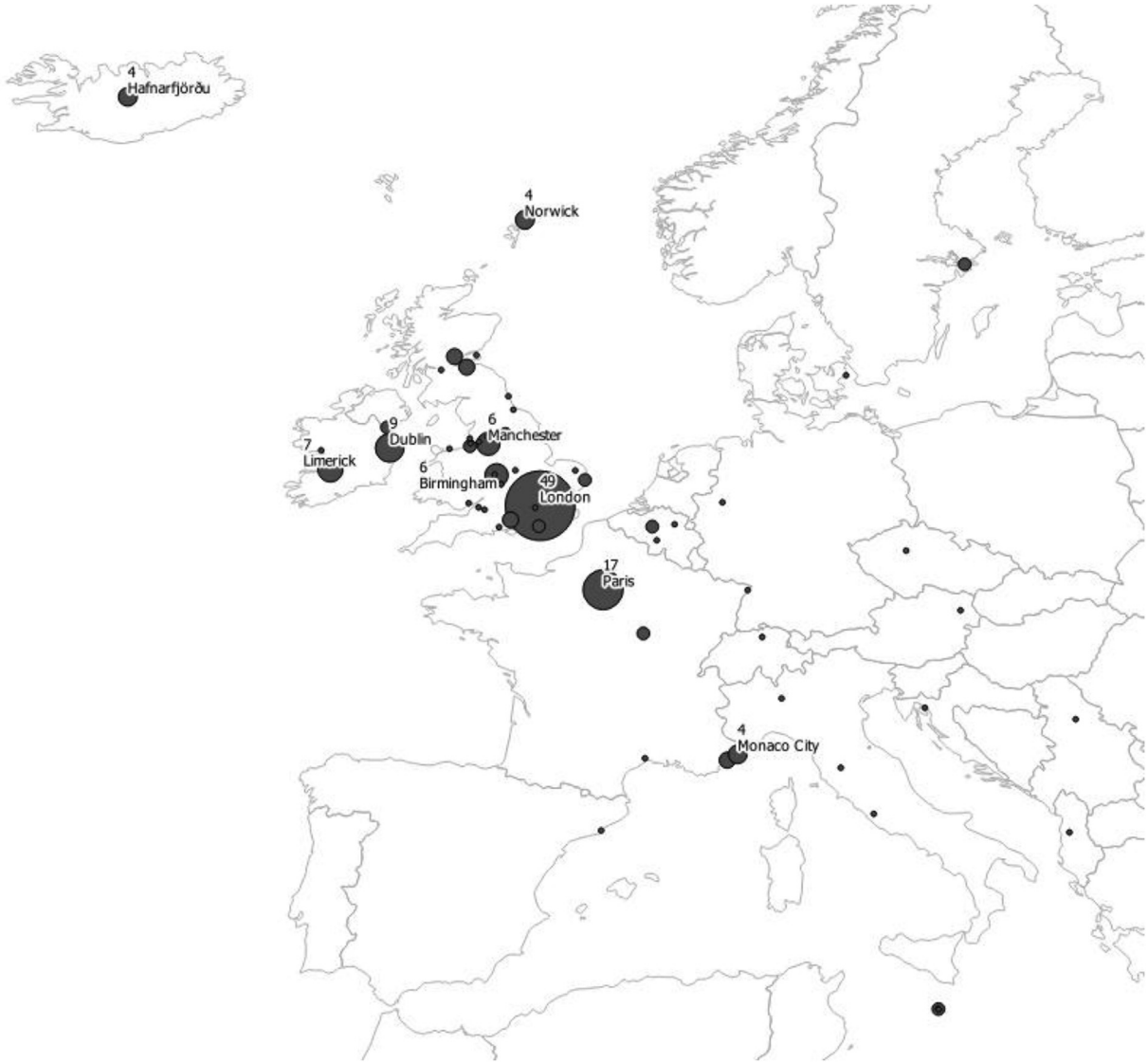

Fig. 5. City-level distribution of jewelry store robberies in Europe (N = 188).

collected using the French keywords. As shown in Figs. 2 and 7, Paris is clearly the city with the highest number of reported robberies in the news. Altogether, Paris, Cannes and Geneva amounted to 20% of the total. In fact, they may concentrate many jewelry stores. The first 10 cities contain more than 30% of the total robberies. Some cities appeared to have a high volume of robberies; the distribution presented on Fig. 2 shows that many regions of France are targeted. No spatial autocorrelation (concentration or dispersion) was detected by the global indicator Moran's I (0,06, $p < 0.001$, queen contiguity with a square grid of 20 km) for the whole distribution. However, Fig. 2 shows some concentration in the south of France as well as in Belgium toward the north of France and in the French-speaking part of Switzerland. A few cases were observed in the west coast of France.

The English dataset shows (Fig. 3) that the main hot spots are the United States, the United Kingdom, Canada, Australia, South Africa and India. The top 20 cities contain more than 30% of total robberies (Fig. 7). London is the most robbed city, followed by Melbourne, New York, Toronto, then Cape Town. Paris appears in both datasets. Fig. 4 shows that the most affected cities in Australia are situated along the east coast, which is the most populated area of the island. Melbourne seems to be the main hot spot. Fig. 5 shows that London is the city that recorded the highest number of jewelry store robberies in Europe, followed by Paris and Dublin. It shows a different pattern than the one observed in Fig. 2 that was produced with the French dataset. The comparison demonstrates the impact of language on the results. In fact, there are major gaps in EMM's incident coverage for some parts of the world.

Fig. 6 demonstrates a strong spatial pattern in North America. Almost all crimes seem to occur on the east and south areas of the United States. New York, Miami, Los Angeles, Houston and Toronto are cities with the highest records. In some regards, the number of covered robberies is correlated to the population of the city. However, some cities with large populations do not have high instances of cases (Chicago, for instance), and some cities with smaller populations present a higher number of cases (Miami, for instance). Thus, the number of cases could be better explained using the number of jewelry stores.

Validity of the country and date of publication as indicators

Table 2 shows how the media of each country covers its own national news. For example, the French media reported 89% of events occurring in France and to a lesser degree news of other countries, such as Monaco, Belgium or Switzerland. Similarly, the United States covered 94% of American events and occasionally covered Canadian, British and French news. These results show that most countries cover events that occur in their own country. The country of publication may thus be used as a proxy to estimate the location of the events in the case of an assessment of risk at the country level.

Finally, the media's rapidity with which it covers events appears excellent. Indeed, 76% of the robberies were covered on the same day of the event in both datasets. About 80% of the events were covered after 10 days, at the latest. As illustrated in Fig. 8, this result confirms that the date of the publication may be used as a valid proxy to describe the period of the event at a macro-temporal analysis level (i.e., an analysis by months, seasons, etc.). In our results, no particular temporal or seasonal trends were detected for jewelry store robberies.

Discussion

The results confirm that online open sources allowed for the detection of jewelry store robberies at the international level. Nevertheless, a comparison with other data sources is still needed to assess the reliability and the ratio of cases covered by online news. Indeed, the verification of the classification was executed to dispose of an errorless dataset ready for further comparison. In the United States, the Jewelers' Security Alliance (JSA) reports 174 and 195 cases in 2016 and 2017, respectively, while the media coverage allowed the identification of an average of 126 robberies per year

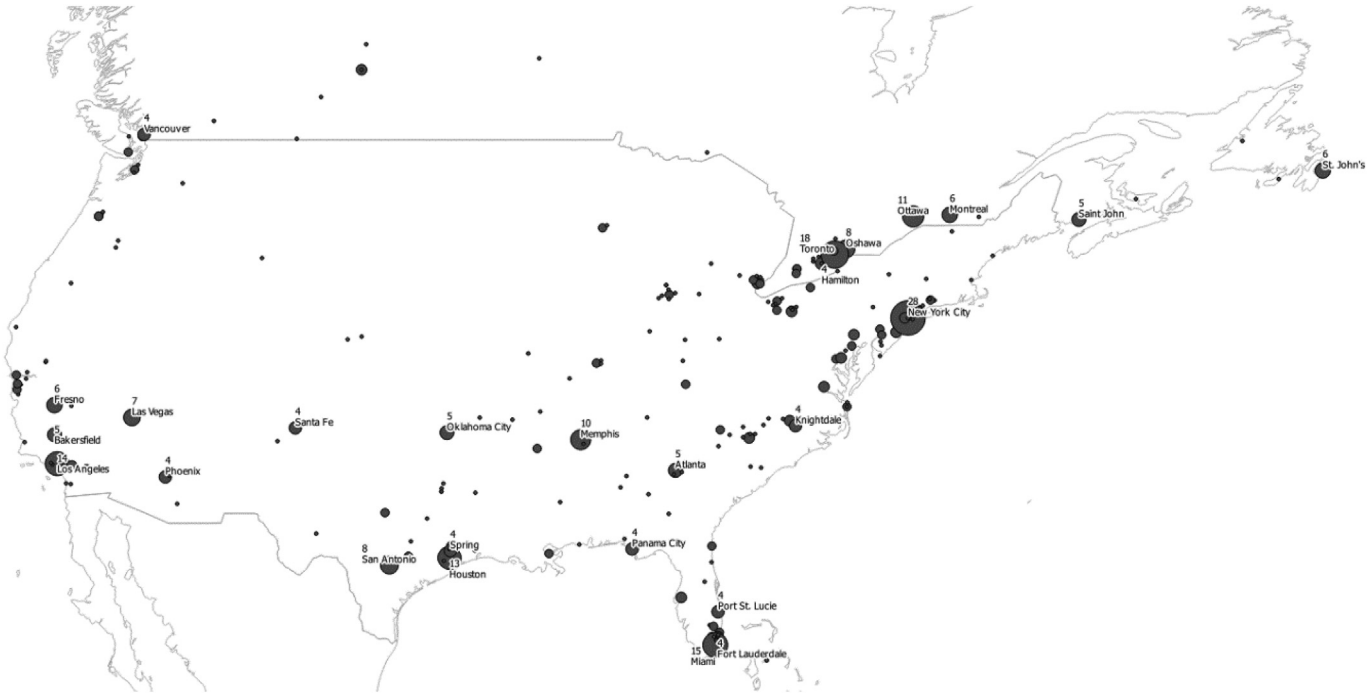

Fig. 6. City-level distribution of jewelry store robberies in the United States and Canada (N = 524).

Table 2
Table showing countries' media coverage according to the French and English datasets. Only Canada appears in both, with a similar proportion of national media coverage (92% in English and 84% in French).

National media coverage	Number of countries	Country names
Exactly 100%	13	Mauritius, Morocco, Algeria, South Africa, Sint Maarten, New Zealand, Trinidad and Tobago, Cyprus, Guyana, United Arab Emirates, Japan, Cayman Islands, Malta
Near 90% and more	5	Australia (94%), U.S. (94%), Canada (92%/84%), India (89%), France (89%)
More than 75%	2	Switzerland (80%), U.K. (77%)
More than 50%	2	Belgium (67%), Reunion (53%)
Fewer than 50%	2	Ireland (46%), Luxembourg (22%)

between 2015 and 2017. Moreover, the spatial pattern in the United States highlighted in Fig. 6 is confirmed by the latest JSA report (Guginsky et al., 2018). This preliminary comparison therefore confirms the hypothesis that the media coverage may be a valid source to evaluate local risks. The findings could also help to detect international cities at risk. It is important to acknowledge that the location of the jewelry store was not considered nor the population density. We inferred that crimes occur in certain areas because they merely reflect where the stores are located. It would therefore be interesting to investigate these hypotheses.

The spatial distribution also shows the great impact of the languages used for the keywords search. The English dataset mainly indicates the English-speaking countries as hot spots, and, in the same way, the French-speaking countries stand out in the French dataset. Almost no Spanish-speaking countries were detected, as expected. Thus a multi-language approach at the data collection stage seems necessary to make a valid comparison among the datasets and draw conclusions independently of the language search. Another option may be the selection of English newspapers for each country of the world (or for those countries one is interested in for analysis). It was observed that the media usually cover the news in their own countries. More rarely, they publish events happening in neighboring countries, where the same language is spoken, or because of the media impact of a particular case. This finding suggests that the country of the publication of a news article can, in principle, be considered a trustworthy indicator of the actual place of the event. Similarly, the date of the publication of the news articles seems to be a good indicator of the actual date of the robbery. In fact, almost 80% of the overall articles reported the event on the same day it occurred. These results show that an early warning system based on filtering relevant news may quickly lead to a global risks assessment at the country level without much data processing efforts.

A second level of analysis needs then to be carried out to detect more specific patterns. It is worth mentioning that deeper knowledge can be obtained from the analysis of the news articles. They

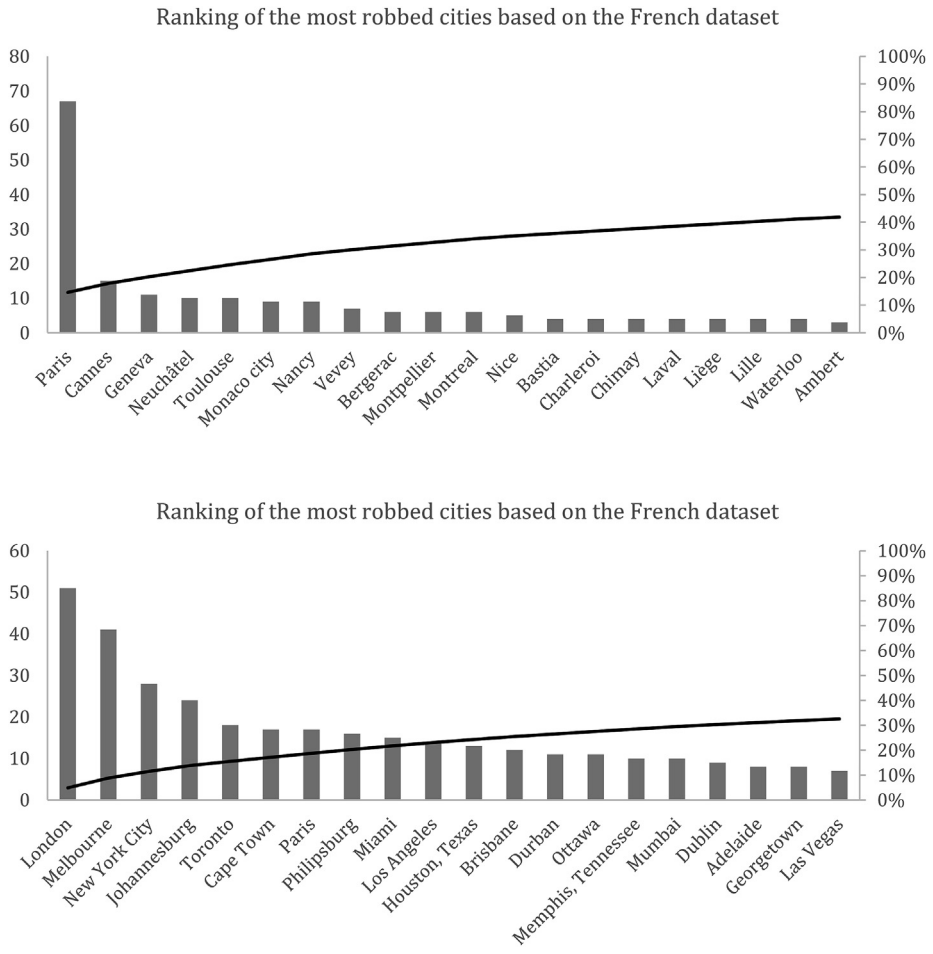

Fig. 7. Pareto diagrams of the top 20 most robbed cities according to the French (top) and English (bottom) datasets.

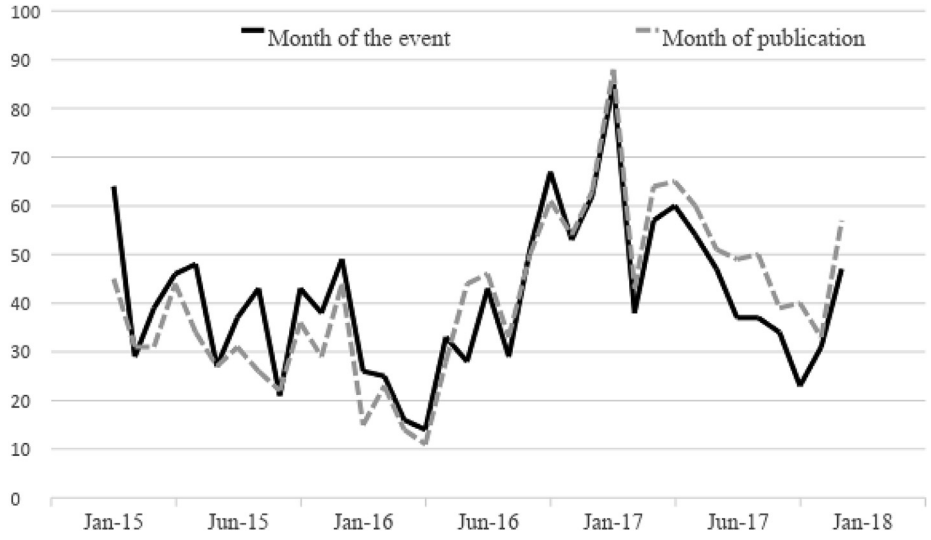

Fig. 8. Comparison of the temporal trends based on the date of the event and the date of publication. Globally, the date of publication appears to be a valid estimator to follow trends at the month level.

often contain information about the geographical area, such as the quarter of the city where the robbery took place, the value and the type of the loot and the modus operandi of the author. However, a dedicated methodology is required to extract the information, for instance, through a named entity extraction approach.

It is clear that what has been discussed in this paper is only a first step to reach the goals set out in the beginning. Indeed, not all the attacks and their severity are the same, and much more information might be extracted from the news articles to generate real intelligence capable of helping jewelry stores design preventative strategies. As mentioned above, information such as the differences among cities, quarters, shops where the robberies take place, the modus operandi of the authors, the value of the loot and the type of weapons used are essential. The next step would therefore be the implementation of algorithms capable of automating the extraction of such entities and their assessment.

Conclusion

This study shows how online news can be used to gather criminal intelligence about a certain phenomenon. Its application to jewelry shop robberies shows that it may be possible to monitor the risks of jewelry store robberies at the international level. To reach this goal, this study focused on the development of a dedicated methodology leading to a working prototype. The results show that open sources are valuable to detect worldwide jewelry store robberies in a fast, simple and low-cost manner. Nevertheless, a cross-check with other data sources, such as police or private security companies, is necessary to assess the validity of the results and therefore the efficiency of the method. This project also shows that it would be possible to extract more accurate information from the cases to allow a better definition of security measures, in regard, for instance, to the modus operandi of the offenders. Much work has to be done with the use of machine learning and natural language processing algorithms to achieve these goals and to better automatize the process. Furthermore, the results show the need to add more languages to the methodology to achieve a more reliable prototype able to fill in the major gaps found in the incident coverage for Latin America, Asia, Eastern Europe and Africa.

Finally, the proposed approach is not limited to jewelry store robberies. In fact, the method developed in this study can most likely be applied to various phenomena that are well covered by the media. It appears to be a transversal method that may be used to evaluate the actual risks and harms of numerous types of security problems. For this reason, it would be interesting to empirically test this hypothesis by applying it to other types of crimes. This would help one assess whether there are phenomena that are more reliably reported than others by the news media and enable one to compare the applicability of this approach to various types of crimes. Online news does not directly trace criminal activities, but it does trace the source of their media coverage. As such, it seems to be a valuable indirect indicator for the analysis of crime trends without having to collect data that are difficult to access, such as police or victim data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aniello, S., Caneppele, S., 2018. Selling stolen goods on the online markets: an explorative study. Glob. Crime 19 (1), 42–62.
Arulanandam, R., Savarimuthu, B., Purvis, M., 2014. Extracting crime information from online newspaper articles. In: Proceedings of the Second Australasian Web Conference, vol. 155. Australian Computer Society, Inc., pp. 31–38
Atkin, H., 2000. Criminal intelligence analysis: a scientific perspective. IALEI Journal 13 (1), 3–7.
Barlow, H., Decker, S.H., 2010. Criminology and Public Policy: Putting Theory to Work. Temple University Press, USA.
Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc.
Chan, E.H., Brewer, T.F., Madoff, L.C., Pollack, M.P., Sonricker, A.L., Keller, M., et al., 2010. Global capacity for emerging infectious disease detection. Proc. Natl. Acad. Sci. 107 (50), 21701–21706.
Chibnall, S., 1977. Law-and-Order News: an Analysis of Crime Reporting in the British Press. Routledge.
Coscia, M., Rios, V., 2012. Knowing where and how criminal organizations operate using web content. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. *ACM*, pp. 1412–1421.
Das, P., Das, A., 2017. Crime analysis against women from online newspaper reports and an approach to apply it in dynamic environment. *Big data Analytics and computational intelligence (ICBDAC)*. In: 2017 International Conference on. IEEE, pp. 312–317.
Dasgupta, T., Naskar, A., Saha, R., et al., 2017. CrimeProfiler: crime information extraction and visualization from news media. In: Proceedings of the

International Conference on Web Intelligence. ACM, New York, NY, USA, pp. 541–549.
Doan, S., Kawazoe, A., Collier, N., 2008. Global health monitor-a web-based system for detecting and mapping infectious diseases. In: Proceedings of the Third International Joint Conference on Natural Language Processing, vol. II.
Ericson, R.V., Baranek, P.M., Chan, J.B., 1991. Representing Order: Crime, Law, and Justice in the News Media. Open University Press, Milton Keynes, pp. 3–4.
Graber, D.A., 1980. Crime News and the Public. Praeger, New York, N.Y.
Guginsky, S.F., Kennedy, J.J., Ruddock, R.O., 2018. 2017 *ANNUAL CRIME REPORT*. 12 March. 6 East 45th Street. Jewelers' Security Alliance, New York, NY 10017.
Jayaweera, I., Sajeewa, C., Liyanage, S., et al., 2015. Crime analytics: analysis of crimes through newspaper articles. Moratuwa Engineering Research Conference (MERCon) 277–282.
Jeni, L.A., Cohn, J.F., Torre, F.D.L., 2013. Facing imbalanced data–recommendations for the use of performance metrics. In: *2013* Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 245–251.
Jewkes, Y., 2015. Media and Crime. Sage.
Marsh, H.L., 1991. A comparative analysis of crime coverage in newspapers in the United States and other countries from 1960–1989: a review of the literature. J. Crim. Justice 19 (1), 67–80.
Mollema, L., Harmsen, I.A., Broekhuizen, E., Clijnk, R., De Melker, H., Paulussen, T., et al., 2015. Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013. J. Med. Internet Res. 17 (5), e128.
Ratcliffe, J.H., 2016. Intelligence-Led Policing. Routledge.
Rohini, D., Isakki, P., 2016. Crime analysis and mapping through online newspapers: a survey. In: Computing Technologies and Intelligent Data Engineering (ICCTIDE), International Conference on. IEEE, pp. 1–4.
Sharma, V., Kulshreshtha, R., Singh, P., et al., 2015. Analyzing newspaper crime reports for identification of safe transit paths. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 17–24.
Taylor, R., 2009. Slain and slandered: a content analysis of the portrayal of femicide in crime news. Homicide Stud. 13 (1), 21–49.