Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour

Ivan Srba, Robert Moro, Jakub Simko, Jakub Sevcech, Daniela Chuda, Pavol Navrat, Maria Bielikova
Slovak University of Technology
Bratislava, Slovakia
{name.surname}@stuba.sk

ABSTRACT

Growing negative consequences of online antisocial behaviour have recently elicited many research efforts, aimed at mitigating or even eliminating this undesired behaviour. However, addressing the open problems is challenging (among other) due to lack of suitable datasets. Also, platforms, where the research results may be applied, are missing too. Therefore, we propose a universal and extensible platform named Monant. It is specifically designed to support characterization and detection of multiple types of antisocial behaviour. Monant does so by means of collecting multimodal, multilingual context-rich data from multiple types of web sources. In addition, the platform supports the deployment of various novel mitigation tools, where data-driven approaches can be applied. To demonstrate the unique characteristics of our platform, we conducted an experimental task in which we monitored healthcare misinformation and identified the most frequent false medical claims related to cancer treatment. Finally, we describe several use cases, which are feasible in our platform and which correspond to trending research directions.

CCS CONCEPTS

• **Information systems** → **Information integration**; *Collaborative and social computing systems and tools*; *Web mining*; *Crowdsourcing*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

antisocial behaviour, misinformation, platform, web monitoring, characterization and detection methods, mitigation tools

1 INTRODUCTION

Antisocial behaviour in online environment is one of the most recent and serious problems. It significantly threatens the principles on which the web was built and also has a critical overreach to society [10]. A typical example is the spread of misinformation via social networks, which influences the opinions and decisions of people (e.g., during shopping, but also when voting or concerning medical treatments). Another typical example is the spread of hate speech or cyberbullying.

ROME 2019, July 25, 2019, Paris, France

Antisocial behaviour has been unintentionally enabled by the rise of information technologies (especially social networking sites, discussion tools and other portals with user generated content). Online anonymity further contributed to its spread. Therefore, researchers worldwide seek to support the process of regulation and elimination of antisocial behaviour through information technologies. A number of approaches addressing various aspects of antisocial behaviour have been proposed so far (for surveys on characterization studies, detection methods, case studies, applications, and fact-checking approaches see [3, 4, 10, 13, 16, 17, 19]).

Data-driven approaches tackling antisocial behaviour fundamentally depend on: 1) access to suitable *datasets*; and 2) *applications* where they can be deployed and applied in practice. While new datasets and applications constantly appear, they do not fully satisfy the needs of current as well as future research of antisocial behaviour. Namely, existing datasets are limited in size, focus on a specific task (e.g., fake news detection), or provide only limited content modalities (most often text) [17]. Similarly, existing applications are implemented as single-purpose tools which are not extensible with new methods and end-user services [3].

To address these limitations, as the part of our ongoing research projects: 1) Automatic Recognition of Antisocial Behaviour in Online Communities (REBELION)[1]; and 2) Misinformation Detection in Healthcare Domain (MISDEED)[2], we propose and implement a platform for monitoring, detection and mitigation of antisocial behaviour named *Monant*. The platform serves primarily as the means for research (to deploy and evaluate new methods and techniques or to conduct case studies) on characterization, detection and mitigation of antisocial behaviour. To a lesser extent, it allows the deployment of research results in practice.

In this paper, we present the proposal and evaluation of the first prototype of our platform. In this already finished prototype, we focused specifically on monitoring and detection of *misinformation and disinformation*. We evaluated its usability in a real-world scenario in which we systematically collected news articles, blogs and fact-checked claims from healthcare domain. The further development and enrichment of the platform with support for additional types of antisocial behaviour (e.g., hate speech detection) represents our future work.

The Monant platform has several unique characteristics, which are the contributions of this work. In contrast to the existing datasets and applications:

[1] https://rebelion.fiit.stuba.sk/
[2] https://misdeed.fiit.stuba.sk/

- It is not focused on one specific task, but it allows research related to *various types of antisocial behaviour* (e.g., fake news, hate speech) and to study interactions between them.
- The collected data are very diverse and rich. It supports *multiple content types* (e.g., news articles, fact-checking articles, blogs, discussions) obtained from *multiple sources* (in terms of a large number of individual sites as well as different source types, such as news portals, discussion tools, social networking sites). It supports *multimodal* (textual, visual or audio) and *multilingual* content. In addition, it allows to consider a *wider context* beyond the content itself (e.g., credibility of authors).
- It is designed to be easily extended by advanced *data-driven methods* that can deal with large amount of unlabeled and dynamically evolving data. It specifically supports interoperability and effective data exchange between various machine-learning-based models (either unsupervised, semi-supervised, supervised or ensemble models). Additionally, it supports active learning as the platform can serve as a mediator between machine learning models and experts. To address dynamically evolving data, it does not provide only historical static data, but it continuously monitors the web and collects data in real time.
- It can also be easily extended by various *novel end-user services*. Examples include an early misinformation warning system or an educational/training tool using real and latest examples of misinformation.

In the rest of the paper, we describe design decisions that enabled these unique characteristics and evaluation of the platform by its employment in practice. Furthermore, we discuss several use cases how the platform and obtained data can open new opportunities for further research.

2 BACKGROUND AND RELATED WORK

At the highest level, antisocial behaviour can be divided into two main categories [9]: 1) spreading of *misinformation or disinformation* (e.g., fake news, fake reviews, rumours, hoaxes); and 2) users' *misbehaviour* (e.g., hating, trolling, manipulation of discussions or cyberbullying). These two categories are, however, closely interconnected. For example, various disinformation is commonly used in manipulation of discussions (so called sockpuppetry) or for the purpose of social spamming by people/bots/cyborgs [17].

Antisocial behaviour is being explored by a permanently increasing body of research. Especially in 2017, the number of publications addressing antisocial behaviour has more than doubled. In general, we identified three major groups of publications, corresponding to three crucial steps of regulation and elimination of antisocial behaviour [10, 17]:

- *Characterization*. The goal of the first group of approaches is to characterize antisocial behaviour by analyzing its manifestations and describing its characteristics. The main subject of characterization is typically an antisocial content and corresponding users responsible for generating this content. To a lesser extent, the context, in which antisocial behaviour occurs, is also analyzed (e.g., especially in case of misinformation, their dynamics and propagation is studied).
- *Detection*. The largest part of the methods aims at detection of antisocial behaviour. Their goal is to automatically or semi-automatically recognize antisocial behaviour. A specifically challenging task is the early detection, i.e., the detection within a short period of time after antisocial behaviour occurred (often with limited information about it).
- *Mitigation* approaches aim to regulate or eliminate antisocial behaviour and its negative consequences. So far, the research dedicated to mitigation has been very limited. There are various strategies to mitigate antisocial behaviour, such as 1) banning/filtering antisocial users and content; 2) use of ranking and selection strategies; 3) education and training to improve human skills in recognition of antisocial behaviour. In case of misinformation, an additional strategy to stop the spread of misinformation is to provide provide facts (coming from an expert or a crowdsourced fact-checking) [3].

Recently (mostly in 2018), a number of valuable survey papers have been published that provide a comprehensive overview of research addressing misinformation in general [3, 10] or focusing on specific forms of antisocial behaviour, such as fake news [13, 16, 17], rumours [19], or hate speech [4].

Summarizing the open problems and challenges from these surveys, we see the following several important research directions in the area of antisocial behaviour.

Exploiting content, user and context data. Currently, the existing characterization and detection methods use only a small part of all available information about the content, the users and their context, specifically mainly textual content with very limited or no context at all. Incorporation of additional content, user and context data can lead to new findings in characterization methods or a higher success rate of detection methods.

Multisource approaches. Most of the current approaches study antisocial behaviour in isolation – typically by using data from one type of source (a news portal, a social networking site, etc.) and even from one particular site. Nevertheless, antisocial behaviour can be detected more effectively by using data across multiple sources (a typical example is fake news detection, where it is possible to consider whether the news articles are offered simultaneously by other reliable/unreliable sources).

Multimodal approaches. Since pictures, movies and audio can be nowadays fabricated and manipulated by means of learning technologies, analyses of multimedia content are essential [10]. Additional modalities, such as visual features [17], should be therefore considered.

Multilingual approaches. The most of existing approaches are restricted to datasets in one language only (English being the most frequent one) [4]. Studies on antisocial behaviour in other languages as well as cross-language approaches are needed.

Extended context. Going beyond the pure content (text or multimedia) represents an additional research potential. For example, credibility of authors and sources can be studied [13]. Readers may provide feedback on antisocial behaviour and thus crowdsourced signals (e.g., reports) should be studied [10]. In case of misinformation, spreading paths and motivations of users/communities to spread misinformation represent another research challenge [3].

Addressing unlabelled and dynamic data. The most of existing approaches are based on supervised machine learning models [17]. Manual labelling of datasets can be, however, very time-consuming and expensive as the process requires careful evaluation of content by experts. In addition, antisocial content (especially in the case of misinformation and disinformation) evolves very dynamically and thus another challenge is to propose methods that can constantly reflect changes occurring in a fast-paced world [16].

Unsupervised, semi-supervised and ensemble models. Unsupervised or semi-supervised models (e.g., co-training) represent an option how to address missing or small datasets. In addition, there is a potential to use various ensemble models.

Active learning. The majority of approaches considers users as passive consumers rather than active co-creators and detectors of antisocial behaviour [3]. In scenarios where only limited or no labelled datasets are available, active learning can be used to systematically incorporate humans and train better models without necessity to have large labelled datasets.

Investigating new mitigation approaches. Besides characterization and detection approaches, mitigation of antisocial behaviour represents many research opportunities as well.

Early warning system. Only few methods proposed so far (e.g., [11]) targets early detection of antisocial behaviour (i.e., detection of antisocial behaviour at its early stages). Nevertheless, the early prediction has a very practical use, since it can significantly contribute to a decrease of the impact of antisocial behaviour [17].

On-site warning system. Currently, detection methods and mitigation tools are many times disconnected from places where people are exposed to antisocial behaviour [3]. These tools should be embedded directly to these environments. So far we can witness only first attempts – in case of misinformation, fact-checkers have been embedded into environments where users consume and share information (usually by means of browser plugins, e.g. [6]).

Education and training. Another option, how to mitigate antisocial behaviour, is to educate people how to recognize antisocial behaviour [10].

Besides these future research directions, the surveys agree that there are no suitable *datasets* to perform this broad spectrum of future research [4, 10, 13, 17, 19]. The existing datasets do not allow standardized comparison of existing methods [10], do not provide multimodal collection of data [13] or do not contain any context [17]. In addition, datasets contain data in one particular language only, and thus there is no place for multilingual research.

Existing *applications* developed to collect data and mitigate antisocial behaviour represent a bottleneck for stated open research problems as well. They are focused on one particular type of antisocial behaviour (mostly misinformation and particularly fake news detection and fact checking) [3]. A typical example present automated fact-checking systems, such as ClaimBuster [7]. Most of the tools and platforms are usually limited to one source of data (e.g., Alethiometer [8] and Fake Tweet Buster [14], which assess content validity on Twitter; Facebook Inspector [2], which aims to detect malicious content on Facebook; or a Chrome browser extension to verify Wikipedia pages [6]). Most of them also consider only text, an exception being an automated assistant proposed by [12] to identify visual news bias. However, it is textual features that are omitted in this case.

Probably the most universal and extensible platform proposed so far, which is also the most similar to our solution, is Hoaxy [15]. It allows to track online misinformation from various sources (social networking sites, fact-checking sites, news sites), detect and visualize misinformation. As the output from the platform, an analysis dashboard is provided. However, it still does not take advantage of multiple types of antisocial behaviour, content multimodality as well as it does not provide more advanced mitigation techniques, to comply with recently identified research directions.

3 MONANT PLATFORM ARCHITECTURE

Our main goal is to address the limitations of the existing antisocial behaviour datasets and applications. To do so, we introduce a platform named *Monant*, which enables the pursuing of the research directions described in Section 2. Monant gathers and stores the data (user generated content with its context) potentially containing antisocial behaviour. The platform then directly integrates the methods used for automated analysis of these data. Results of the methods are also stored to be re-used by further analysis methods or in mitigation scenarios.

The architecture of the proposed solution consists of five high level modules (see Figure 1), which are described in more details in the following subsections: 1) Central data storage; 2) Web monitoring; 3) AI core; 4) Platform management; and 5) End-user services.

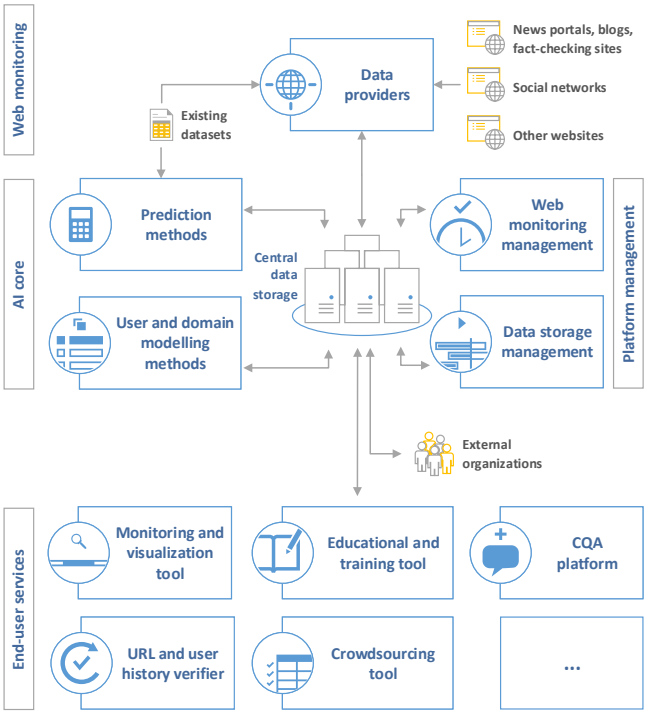

Figure 1: *Monant* platform architecture schema.

So far we have developed the first four platform modules and proposed several end-user services. In future, we plan to continuously improve all modules. Specifically we will focus on research of new methods belonging to the AI core and development of novel

end-user services (please, see Section 5 for possible extensions and supported use cases).

3.1 Central data storage

The central data storage plays a crucial role in the platform. It mediates communication and data transfer between other platform modules. Data storage consists of three layers:

- *Evidence layer* mirrors data as they appear at the source sites and stores them in a unified structure. Evidence layer provides a universal data schema for news articles (including associated multimedia), fact-checking articles and discussions. In addition, it contains additional source-specific schemata (e.g., to store data from YouTube).
- *Inference and prediction layer* contains three types of annotations: 1) content/user features derived from data stored in the evidence layer (e.g., topic model of news articles), 2) machine-learning-based predictions (e.g., whether a news article is predicted to be fake or not) with a corresponding probabilities of prediction being correct, and finally 3) ground-truth labels obtained for example by means of experts. All annotations can describe any entity stored in the platform (e.g., a news article, a source site) or relation between such entities.
- *Platform management layer* contains platform internal data required by web monitoring and data storage management.

From the technological point of view, the central data storage is implemented as an SQL database (PostgreSQL). In order to achieve good interoperability and loose coupling with other platform modules, data access is provided by several groups of REST APIs.

Furthermore, central data storage opens the platform to other organizations and researchers[3]. The goal is to provide our platform to a wider community of researchers and NGOs interested in addressing antisocial behaviour (as a part of our research projects, we have already started a cooperation with several of such stakeholders). The platform allows them to read the already existing content and use this content as part of their own platforms and tools. Alternatively, these external organizations can also take advantage of the AI core in our platform by submitting their own content and getting back the predictions (e.g., to help news publishers with moderation of discussions [18]).

3.2 Web monitoring

The web monitoring module is the main source of data for the platform. It is designed to visit and scrape various kinds of data sources, including news sites, fact-checking sites, social networking sites or various other websites (e.g., information about domain owners). Finally it stores the extracted data to the evidence layer.

Data from these sources are extracted by means of so called *data providers*. Data providers can be used for the purpose of single-shot data extraction as well as for continuous real-time monitoring to identify new antisocial behaviour cases.

The platform supports different types of data providers depending on the structure of the input data. Sites, which do not provide any structured form of data, can be extracted by means of custom web crawlers and parsers. Besides custom site-specific crawlers and parsers, we paid a special attention to data providers which can extract data from a large number of source sites. Since many news, fact-checking sites and blogs provide RSS feeds, similarly as Hoaxy platform [15], we implemented an RSS parser to receive news content. In addition, we employed the `Newspaper` library[4], which can automatically detect the content from news sites. Data from sites, which provide an API (e.g., social networking sites, such as Twitter or YouTube, or news aggregators, such as News API[5]), can be extracted by means of API adapters. Finally, Monant allows additional adapters to import data from the already existing datasets.

A specific feature of our platform is that data providers can be easily configured to be chained together and to exchange data. For example, we can use RSS parser to obtain list of news article URLs and consequently for each new article start a second extraction of article content with the `Newspaper` parser.

From the technological point of view, data providers are implemented in Python. Specifically, `Scrapy` library[6] is used to implement web crawlers, `Beautiful Soup` library[7] to parse HTML content and `feedparser` library[8] to parse RSS feeds.

3.3 AI core

AI core distinguishes our platform from the existing solutions. Conceptually, it is designed as a framework, which allows to easily extend the platform with a wide variety of data-driven methods. AI core aims to enhance the extracted data in the central data storage (evidence layer) with utilization of data analyses, data mining, machine learning (including deep neural networks) and natural language processing.

User and domain modelling methods derive and maintain user and content characteristics, which are stored as features in the inference and prediction layer. These characteristics represent a domain model (e.g., sources and their trust, claims and their validity) and a user model (e.g., authors' credibility, users' previous history).

Prediction methods cover data-driven methods which characterize or detect antisocial behaviour. Their output is stored as predictions in the inference and prediction layer.

As all features, predictions and ground-truth labels used by all methods in AI core are stored at one place, the methods can easily exchange data. For example, as soon as a news article is parsed by any data provider, the first method can model its topics (e.g., by means of Latent Dirichlet Allocation) and store the topic distribution as a feature annotation attached to the corresponding article. The second fake news detection method can consequently take advantage of such pre-calculated feature and use it directly to train/predict article credibility.

3.4 Platform management

The purpose of the platform management is to provide its administrators the necessary tools to manage the data flows between all platform modules.

Web monitoring management. Web monitoring is performed by means of so called *monitors* (an example can be "Monitoring of

[3] In order to obtain access to Monant API, please, contact us by the contact form at: https://rebelion.fiit.stuba.sk/contact

[4] https://github.com/codelucas/newspaper/tree/master/newspaper
[5] https://newsapi.org/
[6] https://scrapy.org/
[7] https://www.crummy.com/software/BeautifulSoup/
[8] https://github.com/kurtmckee/feedparser

health misinformation in Europe"). Each monitor defines which data providers should be used, their scheduling (e.g., frequency of extractions), parameters setup (e.g., a list of RSS feed URLs used as the input to the RSS feed parser) and data provider chaining (if additional data provides should be chained when a new article, discussion, etc. is found). Web monitoring management provides a number of functions to overview extractions, logs and extracted data (see Figure 2 for a screenshot of list and statistics of extracted news articles).

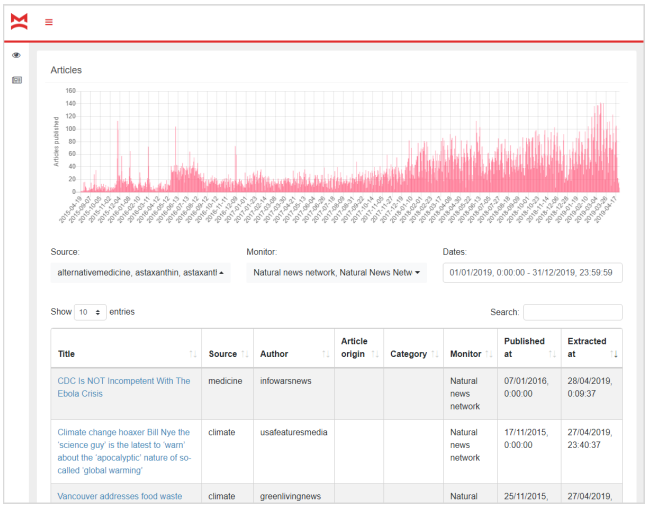

Figure 2: Screenshot from the web monitoring management, which provides a quick overview of the collected data (e.g., distribution of the articles based on their publication date).

Data storage management. The data storage management is primarily responsible for access control to central data storage – user accounts and their access rights. All platform components (e.g., data providers, individual prediction methods, end-user services) are associated with their own account that defines which data can be read, stored, updated or deleted in the central data storage.

3.5 End-user services

Last but not least, the Monant platform enables a wide spectrum of end-user services to be integrated. Their goal is to serve as an interface for experts (e.g., domain-experts such as medical doctors in case of healthcare misinformation, journalist or fact-checkers in case of fake news, etc.) and general public. These end-user services can utilize data from the evidence as well as from the inference and prediction layer of the central data storage. Among them, we list a real-time monitoring and visualization tool or an educational game based on recognizing real true/fake news (determined by verified labels).

4 PLATFORM EVALUATION: MONITORING HEALTHCARE MISINFORMATION

To evaluate the platform design, its implementation and to demonstrate its capabilities, we conducted an experimental task in which we monitored healthcare misinformation and characterized the amount of misinformative articles containing false claims related to cancer treatment. Healthcare represents a domain where misinformation has a critical impact on decision of patients (e.g., refusal of a medical treatment).

For the purpose of this experiment, we implemented custom crawlers and parsers to monitor *Natural News* website[9] together with additional sites belonging to the same network. *Natural News* is a fake news and conspiracy website, which promotes alternative medicine and controversial health claims. Another custom parser and crawler was implemented for a website *Badatel*[10] with similar content in Slovak language (to obtain content in different languages). Additional monitors based on RSS parser, Newspaper crawler and parser and News API parser were configured to cover additional 22 health misinformation sites, which were selected from the list of conspiracy sites maintained by Media Bias/Fact Check[11]. By means of web monitoring management, all monitors were scheduled to run an extraction each hour to achieve continuous access to new appearing articles. This platform module was also used to monitor and verify that extractions were running as expected.

In total we obtained an experimental dataset of 57,959 news articles from 29 sites. Articles were associated with textual content (title, body, tags, categories, references etc.) as well as multimedia (images, videos). Moreover, we extracted information about source sites (country, language) and authors.

Consequently, we created a mapping between obtained news articles and medical claims to evaluate possibilities of AI core. As a source of medical claims we used the existing list of 131 cancer treatments, which are not proved to treat patients, previously prepared in [5]. Each claim in this list is associated with a search query which can be used together with standard information retrieval techniques to obtain relevant documents. The list of claims as well as the presence of claims in news articles were stored in AI core by means of relational annotations between articles and claims.

We were able to map 6,515 news articles (11.2% of all articles) to at least one medical claim (an average number of claims per article is 1.97, a maximal number of claims was 16 in article entitled "Treating Cancer Naturally: 11 Strategies That Work"). The most frequent claims concerning unproved cancer treatments contained mentions of antioxidants (2164 articles), herbalism (1716 articles), Poly-MVA (Lipoic Acid Mineral Complex, 563 articles) and Naturopathy (505 articles).

All data (news articles, claims, relation annotations capturing mapping between claims and articles) were stored in the central data storage and thus they can be simply accessed by other researchers, organizations or utilized by end-user services.

The carried-out evaluation demonstrates the platform capabilities in data collection and their annotation by the means of claims. In the future, we plan to extend the list of medical claims (e.g., from fact-checking sites, such as *Snopes*[12]) to be able to map more articles and thus to characterize the spread of medical misinformation.

[9] https://naturalnews.com/
[10] https://www.badatel.net/
[11] https://mediabiasfactcheck.com
[12] https://www.snopes.com/

5 SUPPORTED USE CASES

We provide several use cases, which are enabled either by data, which can be collected by Monant platform, or by its overall architecture and design. They demonstrate the contributions and unique characteristics of the platform stated in Section 1 and highlight how Monant supports the research directions stated in Section 2.

Taking advantage of various antisocial behaviour types. While we focused so far on misinformation and disinformation, the proposed platform is not restricted to one particular antisocial behaviour type. We can use the same collected data to study various antisocial behaviour types (e.g., detect fake news in news articles and hate speech in the attached discussions) as well as to study interactions between them since a variety of antisocial content is many times created, discussed or shared by misbehaving users (e.g., bots, social spammers).

Exploiting content, user and context data. When characterizing or detecting the antisocial behaviour, a related content (even of a different type) may be useful. For example when detecting fake news, the attached discussions can reveal features that may not be possible to derive from the news' content itself. Conversely, when detecting hate speech in a discussion, the attached news/multimedia content can provide new features.

We can take an advantage of a wide spectrum of monitored sources. For example, news articles can be grouped together by referencing the same news story. Consequently a whole new feature set can be derived from analyzing other news in the same story, such as the proportion of fake news, the closest (textual) similarity with the fake/true news.

Since our collected data contain all modalities of content (text, images, etc.), we can consider all these modalities during the feature engineering process. For example, we can track the source of image attached to news article and detect if it is used by other reliable/unreliable sources.

If the content is in different languages, transfer learning can be used to detect antisocial behaviour for languages with no or small labelled datasets. Finally, through user and domain models, we can build new contextual features such as author or source credibility.

Addressing unlabelled and dynamic data. The platform allows us to combine human expertise with machine learning through active learning. When the prediction methods are not sure about their prediction (e.g., when a new misinformation case emerges), experts can be asked to provide the correct labels by means of a crowdsourcing tool or a CQA platform. Afterwards, the obtained labels will be incorporated into the central data storage and thus, be prepared to be used in the next iteration of model training.

Investigating new mitigation approaches. New mitigation approaches can be researched by means of end-user services, e.g.:

- *Monitoring and visualization tool.* It will allow users (journalists, NGOs, but also general public) to monitor and visualize in real-time the amount of antisocial behaviour appearing in different sources.
- *URL and user history verifier.* It will allow users to verify, whether the provided URL contains misinformation, hate speech or other kinds of antisocial behaviour. The extended version of this tool can automatically analyze users' profiles on social networking sites (e.g., shared URLs at Twitter or Facebook) and provide overall profile statistics (e.g., the proportion of URLs containing misinformation).
- *Education and training tool.* We can take advantage of the real-world labelled data and use them to train users in better detection of antisocial behaviour. For example, Cohen et al. [1] designed a chatbot for simulated conversation which purposefully contains the displays of cyberbullying.
- *Crowdsouring tool.* To reduce the problem with time-consuming manual data labelling, we can employ the power of crowd. Gamification can be used to engage users even more.
- *CQA platform.* CQA (Community Question Answering) platform can serve to mediate the communication about the misinformation. It can connect experts, non-expert users and even prediction methods from AI core. AI core methods could create questions (fact-checking requests as a part of active learning) as well as provide answers (if methods can predict article credibility with sufficiently high probability).

6 CONCLUSION AND FUTURE WORK

The research on antisocial behaviour has significantly increased in recent two years (starting from 2017). New forms of antisocial behaviour constantly emerge. Due to this dynamism, many open problems and research challenges remain unanswered. One of the inhibitors of the research progress is the lack of rich and standardized data. Such data would enable the proposal of new methods and comparison of the performance of the existing ones.

In order to relieve the high demand for richer datasets, we propose a universal and easily extensible platform Monant for monitoring, detection and mitigation of antisocial behaviour. In contrast to existing applications, it supports multiple types of antisocial behaviour and content/context-rich data. In addition, Monant serves as a framework which allows to easily plug in various data-driven methods for characterization and detection of antisocial behaviour and to deploy novel end-user mitigation services. While the platform is a subject of a long-term development, the first prototype of crucial parts of this platform is already developed and deployed and we have successfully confirmed its viability.

In our future work, we plan to enhance the web monitoring module with additional data providers. Furthermore, the prediction methods and user/domain modelling methods will be the primary subjects of our research efforts. Last but not least, we plan to develop some end-user services to make the results of our research accessible to expert users as well as general public. Moreover, we plan to open the platform API, its AI core as well as the obtained data to all researchers and other stakeholders, who are interested in the study of antisocial behaviour.

ACKNOWLEDGMENTS

This work was partially supported by the Slovak Research and Development Agency under the contracts No. APVV-17-0267, APVV SK-IL-RD-18-0004 and by the Scientific Grant Agency of the Slovak Republic, under the contracts No. VG 1/0725/19 and VG 1/0667/18. The authors would like to thank the students, who contributed to design and implementation of the first prototype of the Monant platform, and the STU Grant scheme for Support of Excellent Teams of Young Researchers.

REFERENCES

[1] R Cohen, Z. Liao, A. Mancisidor, S. Nagpal, A. Pham, A. Saini, J. Shen, H. Singh, C. Tavares, S. Thandra, N Mathiarasu, R Aarif, S. Ansari, D Fraser, M Hegde, J. Henderson, I. Kajic, and A. Khan. 2018. An education-based approach to aid in the prevention of cyberbullying. *ACM SIGCAS Computers and Society* 47, 4 (jul 2018), 17–28. https://doi.org/10.1145/3243141.3243146

[2] Prateek Dewan and Ponnurangam Kumaraguru. 2017. Facebook Inspector (FbI): Towards automatic real-time detection of malicious content on Facebook. *Social Network Analysis and Mining* 7, 1 (dec 2017), 15. https://doi.org/10.1007/s13278-017-0434-5

[3] Miriam Fernandez and Harith Alani. 2018. Online Misinformation: Challenges and Future Directions. In *Comp. of the The Web Conf. 2018 on The Web Conf. 2018 - WWW '18*. ACM Press, New York, New York, USA, 595–602. https://doi.org/10.1145/3184558.3188730

[4] P. Fortuna and S. Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *Comput. Surveys* 51, 4 (jul 2018), 1–30. https://doi.org/10.1145/3232676

[5] Amira Ghenai and Yelena Mejova. 2018. Fake Cures: User-centric Modeling of Health Misinformation in Social Media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (nov 2018), 1–20. https://doi.org/10.1145/3274327 arXiv:1809.00557v1

[6] Reed H Harder, Alfredo Velasco, Michael Evans, Chuankai An, and Daniel Rockmore. 2017. Wikipedia Verification Check: A Chrome Browser Extension. In *Proc. of the 26th Int. Conf. on World Wide Web Comp. - WWW '17 Companion*. ACM Press, New York, New York, USA, 1619–1625. https://doi.org/10.1145/3041021.3053364

[7] Naeemul Hassan, Anil Kumar Nayak, Vikas Sable, Chengkai Li, Mark Tremayne, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, and Aaditya Kulkarni. 2017. ClaimBuster: The First-ever End-to-end Fact-checking System. In *Proc. of the VLDB Endowment*, Vol. 10. VLDB Endowment, 1945–1948. https://doi.org/10.14778/3137765.3137815

[8] Eva Jaho, Efstratios Tzoannos, Aris Papadopoulos, and Nikos Sarris. 2014. Alethiometer: a Framework for Assessing Trustworthiness and Content Validity in Social Media. In *Proc. of the 23rd Int. Conf. on World Wide Web - WWW '14 Companion*. ACM Press, New York, New York, USA, 749–752. https://doi.org/10.1145/2567948.2579324

[9] Srijan Kumar, Justin Cheng, and Jure Leskovec. 2017. Antisocial Behavior on the Web: Characterization and Detection. In *Proc. of the 26th Int. Conf. on World Wide Web Companion*. Int. World Wide Web Conferences Steering Committee, 947–950. https://doi.org/10.1145/3041021.3051106

[10] Srijan Kumar and Neil Shah. 2018. False Information on Web and Social Media: A Survey. In *Social Media Analytics: Advances and Applications*. CRC press.

[11] Srijan Kumar, Francesca Spezzano, and V.S. Subrahmanian. 2015. VEWS: A Wikipedia Vandal Early Warning System. In *Proc. of the 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining - KDD '15*. ACM Press, New York, New York, USA, 607–616. https://doi.org/10.1145/2783258.2783367

[12] Vishwajeet Narwal, Mohamed Hashim Salih, Jose Angel Lopez, Angel Ortega, John O'Donovan, Tobias Höllerer, and Saiph Savage. 2017. Automated Assistants to Identify and Prompt Action on Visual News Bias. In *Proc. of the 2017 CHI Conf. Ext. Abstracts on Human Factors in Computing Systems - CHI EA '17*. ACM Press, New York, New York, USA, 2796–2801. https://doi.org/10.1145/3027063.3053227

[13] Shivam B. Parikh and Pradeep K. Atrey. 2018. Media-Rich Fake News Detection: A Survey. In *Proc. of IEEE Conf. on Multimedia Information Processing and Retrieval - MIPR '18*. IEEE, 436–441. https://doi.org/10.1109/MIPR.2018.00093

[14] Diego Saez-Trumper. 2014. Fake Tweet Buster: A Webtool to Identify Users Promoting Fake News on Twitter. In *Proc. of the 25th ACM Conf. on Hypertext and social media - HT '14*. ACM Press, New York, New York, USA, 316–317. https://doi.org/10.1145/2631775.2631786

[15] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A Platform for Tracking Online Misinformation. In *Proc. of the 25th Int. Conf. Comp. on World Wide Web - WWW '16 Companion*. ACM Press, New York, New York, USA, 745–750. https://doi.org/10.1145/2872518.2890098

[16] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 3 (2019), 21. https://doi.org/10.1145/3305260

[17] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (sep 2017), 22–36. https://doi.org/10.1145/3137597.3137600

[18] Andrej Švec, Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2018. Improving Moderation of Online Discussions via Interpretable Neural Models. In *Proc. of the Second Workshop on Abusive Language Online - ALW '18 at EMNLP*. Association for Computational Linguistics, 60–65. http://aclweb.org/anthology/W18-5108

[19] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media. *Comput. Surveys* 51, 2 (feb 2018), 1–36. https://doi.org/10.1145/3161603