ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240460

Back to the Future: Symbolic Reasoning to Combat the Malicious Use of Social Media

Maria Vanina Martinez

Artificial Intelligence Research Institute (IIIA-CSIC), Spain
ORCID (Maria Vanina Martinez): https://orcid.org/0000-0003-2819-4735

Abstract. As technology widens our possibilities for communication and knowledge discovery, it can also leave us vulnerable to misinformation and toxic, abusive, manipulative content, which can cause substantial harm both individually and collectively. Combating the malicious use of social media implies solving very complex and subjective tasks that require a synergy between automated tools and humans. Existing ML models, trained on vast datasets, excel in many tasks like summarization, translation, and human-like content generation. However, we argue that in order for those tools to be effective, they need to be able to form the types of representations and semantic inferences that humans use and create, have the ability to combine automatically acquired data and insights with the domain-specific expertise of the users, and involve high-level reasoning, all three features in which knowledge-based and symbolic approaches to artificial intelligence seem to be well-fitted for.

1 Introduction

The last two decades of technological progress have produced an increasingly digitalized and interconnected society. Different types of computer systems coexist with human beings enabling efficient communication, access to information, and helping carry out a wide range of daily activities. We call these systems, socio-technical [26], as they intervene, directly or indirectly, in activities related to a variety of social aspects. Social (media) platforms are a clear example of socio-technical systems used seamlessly throughout our daily lives. People interact constantly on these platforms and increasingly base their decisions on the content they consume. Although undoubtedly useful, they also expose users to different types of cyber attacks such as cyber bullying, grooming, and hate speech, among others, that have the potential for causing great damage, both individually and socially [27, 43], and pose significant challenges to security and trust.

The tasks of identification, analysis, and monitoring of this type of phenomena, have received great interest from the scientific community in recent years, with some advances in Artificial Intelligence (AI), mostly by means of Machine Learning (ML) tools. Social media companies invest huge amounts of resources in cybersecurity[1] in order to try to stop this kind of abuses [1]. However, the complexity of the analysis is beyond the capabilities of the tools provided by the state of the art, largely due to the subjectivity required to understand the content, which in most cases requires prior knowledge of the domain, contextual information, and common sense reasoning. These phenomena are multidimensional, that is, there are a variety of aspects that characterize them and that impart different types of risks to users. It is clear then, that these processes need software tools that can help people (e.g., analysts) identify and manage malicious content (and intent) in social media in an effective and sustainable way [2].

In this paper, we argue that in order to effectively tackle such issues, it is necessary to design tools that integrate knowledge and reasoning with learning. We show a series of efforts towards this goal through the formalization and construction of hybrid AI tools that combine learning with logic-based reasoning and knowledge representation models. The overarching line of research that I conduct aims to deepen the study and development of knowledge-based AI models that, combined with state-of-the-art ML methods, make it possible to formulate systems that, on one hand, help researchers and analysts better understand the intricate reality of social platforms and also, help users navigate the spectrum of social networks safely. These systems must be capable of high-level reasoning, which implies, at a minimum, having the ability to combine data and automatically acquired knowledge with expertise in a specific domain, and also being able to construct and understand the types of semantic representations and inferences that humans use and create. Our hypothesis is that knowledge-based models, in particular those based on logic, can be helpful in providing such capabilities, especially when focusing on identifying and combating malicious use of social media.

The paper is organized as follows. In Section 2 we discuss work on Hate Speech and the incorporation of argumentation theory in the automatic generation of counternarratives. In Section 3 we describe several approaches to interactions in social platforms. Finally, in Section 4 we describe a framework to develop hybrid intelligent socio-technical systems, and we show its application in the study of several tasks related to monitoring and combating the malicious use of social media, more specifically, adversarial deduplication, fake news spread and detection, and monitoring of social media content.

2 Hate Speech and Counternarratives

Hate speech has accompanied human evolution, however, by means of social media and new technologies such as Generative AI, hate speech can be amplified beyond human scale, spreading faster and increasing its reach, furthermore, coupled with disinformation it can

[1] Cybersecurity can be informally defined as "the protection of systems (including software, hardware, or humans) connected to the internet".

[2] One of the problems with human moderators is that they are often exposed to toxic content for long workdays.

lead to stigmatization, discrimination and large-scale violence[3][4].

The most predominant strategy adopted so far by social media companies to counter hate speech is to recognize, block, and delete these messages and/or the users that generated it. This strategy has two main disadvantages: first, though it prevents a message from spreading further, blocking and deleting does not counter its consequences on those who were already exposed to it. Second, since the response is of binary nature, so is the classification of a message, i.e., a message is or is not hate speech, leaving no place for subtleties or shades of interpretation. This strategy seems to be an overly simplistic approach to deal with such an inherent complex phenomenon and can generate accusations of overblocking or censorship.

An alternative that has gained attention in the last years, is to oppose hate content by responding directly to the message, refuting or undermining it [4, 7][5]. In this way, the consequences of mistakes in speech classification are minimized, overblocking is avoided, and it helps to spread a narrative against hate that can reach people that are not necessarily convinced, or not involved in the conversation[6].

The huge volume of online messages makes it clear that such actions cannot be carried out effectively in a "manual way", even with thousands of volunteers taking part in this effort independently or in coordination. In this scenario, automated generation of counter-narratives is a tempting avenue; however, apart from being a very complex and subjective task of natural language interpretation, this task also poses a great challenge due to the complex linguistic and communicative patterns involved in argumentation. Though traditional ML approaches have typically produced less than satisfactory results for argumentation mining and generation, the recent availability of Large Language Models (LLMs) provides a promising approach to address such tasks in general and of counter-narratives generation in particular. Recent work shows that LLMs are capable of tackling several argumentative reasoning tasks with some degree of success [40]. However, as [23] show in their in-depth analysis of the argumentative capabilities of GPT-3, although the language they generate is clearly argumentative, most of them are not considered acceptable arguments by humans, falling in fallacies like "begging the question" and providing mostly irrelevant information.

In our line of work related to hate speech, we argue that the identification of argumentative information within the messages can improve the quality of arguments generated by LLMs, more concretely, in improving the quality of automatically generated counter-narratives against hate speech. In [14] we studied this hypothesis by comparing the following scenarios: (i) using LLMs without any specific adaptation to the task or domain, (ii) using LLMs that have been fine-tuned using a dataset of counter-narratives, (iii) using LLMs in a few-shot approach, and (iv) providing to the model additional information about some of the argumentative aspects of the hate speech.

The datasets used for training, fine-tuning, few-shot, and testing were developed in [13] and [14]. In [13], the dataset ASOHMO (Argumentation Structure Of Hate Messages Online), was created by enriching the Hateval corpus [3] with a manual annotation of their argumentative aspects, adapted from [44]'s Periodic Table of Arguments, an analytic approach to represent the semantics of the core argumentative schemes proposed in [45], but with fewer categories and based on a limited set of general argument features; it contains tweets both in English and Spanish. We identified the following argumentative aspects of hate speech in tweets:

- **Justifications** and **Conclusions**.
- **Type** of Justification and Conclusion: Fact, Policy or Value.
- A **Pivot** signalling the argumentative relation between Justification and Premise.
- Domain-specific components: the **Collective** which is the target of hate, and the **Property** that is assigned to such Collective.

In [14], we presented CONEAS (Counter-Narratives Exploiting Argumentative Structure), a dataset of counter-narratives defined according to the argumentative information labeled on tweets from ASOHMO [13]. Each argumentative tweet is paired with counter-narratives of three different types defined by applying systematic transformations over argumentative components of the tweet, and a fourth type consisting of any counter-narrative that does not fall under any of the other three [7].

The types of counter-narrative we utilize are the following:

Negate Relation Between Justification And Conclusion Negate the implied relation between the justification and the conclusion.

Negate association between Collective and Property Attack the relation between the property, action, or consequence that is being assigned to the targeted group, and the targeted group itself.

Attack Justification based on its type If the justification is a "fact", then the fact must be put into question or sources must be asked to prove that fact. If it is of type "value", it must be highlighted that the premise is actually an opinion. If it is a "policy", a counter policy must be provided.

Free Counter-Narrative If the annotator recognizes a counter-narrative that does not fit in any of the previous types they are encouraged to write it down under this fourth type.

TWEET: `@user must deport all illegal migrants india already reeling under constant threat of muslim radicals curb population`

Justification: `india already reeling under constant threat of muslim radicals curb population` **(fact)**

Conclusion: `must deport all illegal migrants` **(policy)**

Collective: `illegal migrants`

Property: `muslim radicals`

Negate relation between justification and conclusion: *Deporting illegal migrants will not mitigate the problems with muslim radicals.*

Negate relation between collective and property: *Illegal migrants are not necessarily muslim radicals.*

Negate justification based on type: *It is not true that India is reeling under threat of muslim radicals.*

FREE COUNTER NARRATIVE *Deporting illegal migrants without consideration to their circumstances is inhumane.*

Figure 1: Examples of each type of counter narratives [14].

[3] https://www.un.org/en/hate-speech/understanding-hate-speech/hate-speech-and-real-harm

[4] https://www.unesco.org/en/articles/new-unesco-report-warns-generative-ai-threatens-holocaust-memory

[5] http://www.nohatespeechmovement.org/

[6] *The Dangerous Speech Project* literature review on the effectiveness of counter-speech actions - https://dangerousspeech.org/wp-content/uploads/2021/06/Counterspeech-annotated-bib-as-published-2020.docx.pdf

[7] All counter-narratives, regardless of their type, follow the guidelines of the Get The Trolls Out project - https://getthetrollsout.org/stoppinghate

Automated counter-narrative generation has been recently tackled by leveraging the rapid advances in neural natural language generation. As with most natural language generation tasks in recent years, the basic ML approach has been to train or fine-tune a generative neural network with examples specific to the target task. Examples of these are [7, 33, 46, 38, 12, 5, 9, 2, 42, 41]. However, none of the aforementioned datasets or approaches to counter-narrative generation includes or integrates any additional annotated information apart from the hate message, possibly its context, and its response. These datasets are well-suited for neural sequence to sequence approaches [34, 37], which take one text string and output another. In this case, they take a hate narrative and output a counter-narrative. That is why we consider an alternative approach that aims to reach generalization not by just providing a huge number of examples (that may not always exist depending on the language or the topic), but by providing a richer analysis of such examples that guides the model in finding adequate generalizations. We believe that information about the argumentative structure of hate speech, may be used as constraints for automatic counter-narrative generation. Apart from generating counternarratives, in [13] we showed that our approach allows also for the identification of such argumentative structures and components, which could provide analysts with a basis for explanations once the message has been classified as hate speech. Most aligned to our approach is [8], where they address an argumentative aspect of hate speech countering. They classify counter-narratives by type, using a LLM, and show that knowledge about the type of counter-narratives can be transferred across languages, but they do not use this information to generate counter-narratives.

The evaluation of counter-narratives is not straightforward either. So far, no automatic technique has been found satisfactory for this specific purpose. Automatic metrics proposed for other NLP tasks, like BLEU [28] for automatic translation or ROUGE [25] for summarization, are not adequate as they strongly rely on word or n-gram overlap with manually generated examples. These measures are disputed in the NLP community, mostly regarding NL generation, because they cannot be adapted to cases where there can be many possible good outputs of the model, with significant differences between themselves. In our case, valid counter-narratives can present strong variations not only in the words used but even in the semantics and communicative intentions that underlie a given choice of words.

Without adequate automatic metrics, many authors have resorted to manual evaluations for automatically generated counter-narratives. Such evaluations often distinguish different aspects of the adequacy of a given text as a counter-narrative for another. In [9], they evaluate three aspects of the adequacy of counter-narratives: *Suitableness* (if the counter-narrative was suited as a response to the original hate message), *Informativeness* (how specific or generic the response is) and *Intra-coherence* (internal coherence of the counter-narrative regardless of the message it is responding to). In [2], they assess these three other aspects: *Offensiveness*, *Stance* (towards the original tweet) and *Informativeness* (same as [9]). Based on these previous works, in [14] we proposed initial criteria to manually evaluate the adequacy of counter-narratives, considering four different aspects:

- **Offensiveness**: if the tweet is offensive to either the target group, the author of the tweet, or any other group or person. Possible values are: Offensive; Possibly Offensive/Not clear; Not offensive.
- **Stance**: if the tweet supports or counters the specific message of the hate tweet. Possible values are: Supports the original message; Not clear/Changes subject wrt original tweet; Counters the original message.
- **Informativeness**: Evaluates the complexity and specificity of the generated text. Only counter-narratives with a "Counters" Stance are evaluated. Possible values are:
  1. **Generic statement**: replies that do not incorporate any information mentioned on the tweet and could counter many different hate messages, e.g, "I don't think so" or "That is not true".
  2. **Specific but not argumentative**: the reply is a simple statement, possibly composed of a single sentence without providing justification for the stance, e.g., sentences like "I don't think that" or "Do you have proof that" followed by a verbatim copy of some part of the hate tweet.
  3. **Specific and Argumentative**: counter-narratives with some degree of elaboration of the information contained in the hate message. We identified three common patterns:
     - A - replies that take more than one element from the original message and establish some relation between them (e.g. "I don't see the relation between {*element from the original message*} and {*other element from the original message*}").
     - B - A statement declaring stance over an element from the original tweet but adding a second coordinated statement with personal appreciations about it (e.g. "I don't think we should {*some policy mentioned on the tweet*}. It is a bad idea").
     - C - An argumentative reply based on information not mentioned explicitly in the original tweet, but necessarily inferred, showing a comprehensive understanding of the meaning of the hate message (e.g. a reply to a tweet concluding with #BuildTheWall saying "*Building a wall would cost the taxpayers more*" or "*Building a wall won't give you more control over illegal trafficking*").
- **Felicity**: Evaluates if the generated text sounds, by itself, fluent and correct [8]. The possible values are: The text is incoherent or semantic or syntactically incorrect; The text is coherent with small errors like incoordination of genre/tense/etc. or repeating parts of the original text without adapting them to the text being generated; The text is fluent and sounds correct.

To aggregate the results for these four categories, we define two extra concepts: *Good* and *Excellent* counter-narratives. Good counter-narratives are those with optimal values on Offensiveness, Stance, and Felicity. Excellent counter-narratives also have the optimal value of informativeness. We believe Informativeness is the most valuable of the four categories, which is why it is determinant in characterizing Excellent counter-narratives. The Good indicator shows that productions are not harmful or totally random.

To assess the quality of the counter-narratives generated in the different scenarios, we carried out a preliminary evaluation with human evaluators, who achieved moderate agreement with each other. Based on those judgments, we can say that argumentative information by itself does not produce an improvement in the counter-narratives, but high-quality, specifically targeted fine-tuning seems to have a positive impact. Argumentative information does produce improvements in scenarios with very small training data and very specific fine-tuning, which seems promising to produce highly tailored counter-narratives, as in Gupta et al. [22]. We are at the moment conducting a broader experiment, in which we have incorporated some of the latest LLMs for evaluation.

[8] It is related to [9]'s Intra-Coherence, but also considers additional dimensions like syntactical and semantic correctness.

3 Modeling Social Media Interactions

It is undebatable that social interactions mold, to different degrees, people's knowledge, beliefs, personalities, and behavior. The formalization of these interactions plays an important role in the design of cybersecurity tools. In [18], we focused on reasoning about the diffusion of beliefs in social media. A Network Knowledge Base (NKB), models sources of social communication among users, who also have access to a stream of news items that are produced by others. NKBs adopt the typical model of social networks as sets of agents with various relationships among them and, each agent (a user in the social network) has its own knowledge base (KB). Each agent is represented as a vertex, and relationships are represented as arcs between the vertices. NKBs can thus be seen as complex multilayer networks that allow representing the individual beliefs of each network node, as well as multiple attributes of the nodes and their relationships, affording the possibility of combining models for more than one social platform. Feeds—the pieces of information that each user sees when engaging with platforms—are modeled as news items (sentences in a logical language) that represent the source, content, and an indication of whether the user who posted it is signaling an addition or deletion to their own KB.

In [18], we defined and formalized local revisions, i.e., responses/modifications to the local knowledge bases on their feeds), and each local revision is carried out in parallel. In [16] we report preliminary experiments on Twitter data showing that different agent types react differently to the same information. In [17] we processed raw data obtained from social media based on the framework defined in [18] and [19], and then formulated an action/no action prediction task that takes as input five features that include the user's personality type among other social cues. We showed via an extensive empirical evaluation with real-world Twitter data that machine learning classification algorithms can be successfully applied in this setting to make simple predictions about user reactions. However, recent efforts towards generalizing such models to predict not only the intention of a user of responding to a message, but also potential values for several linguistic features relevant to the ethos and pathos of online discussion [15], showed that ML models alone may not be sufficient, and we are experimenting with hybrid models that integrate ML with classical rule mining as well as expert knowledge. This is part of ongoing collaborations within the framework of the *iTRUST* project, seeking the understanding polarization on social media and its effects [9].

4 An Architecture for Hybrid Intelligent Systems

In this section, we show a more general perspective on the problems related to the malicious use of social media.

From a cybersecurity point of view, many of the problems associated with the malicious use of social media are not necessarily independent of each other and some of them could be addressed using similar tools or methodologies. We argue that these problems, and many others, can be effectively addressed by (i) combining multiple data sources that are constantly being updated, (ii) maintaining a knowledge base using logic-based formalisms capable of *value invention* to support generating hypotheses based on available data, and (iii) maintaining a related knowledge base with information regarding how actors are connected, and how information flows across their network. For this reason, in [30, 39], we proposed HEIST (Hybrid Explainable and Interpretable Socio-technical Systems), which is an application framework [10] that aims to guide the implementation of hybrid socio-technical systems that require explainable outputs. The motivation for defining such architecture follows the idea that in order to build truly intelligent and trustworthy socio-technical systems, their behavior needs to be based on knowledge, not just domain knowledge but also related to the user and how the system is used.

We briefly describe each of the six components, referring the reader to [39] for a full description. For an illustration of the architecture, see Figure 2.

Data Ingestion: Handles the integration of data sources, addressing basic issues like data cleaning, schema matching, inconsistency, and incompleteness management. It also deals with higher-level challenges such as trust and uncertainty management, ensuring the proper handling of heterogeneous data.

Subsymbolic Services: This module focuses on tasks that are best solved using data-driven (e.g., ML) services. Having such tools isolated in a module helps to identify specific application scenarios for each service, facilitating faster implementation, providing alternative implementations, and the generation of explanations.

Symbolic Reasoning: High-level reasoning is key for addressing general problems. This module, which serves as the core of the framework, leverages preprocessed data from the Data Ingestion Module and outputs from the Subsymbolic Services Module. Rule-based systems are commonly employed here to perform complex tasks like combining low-level data and knowledge or providing responses based on well-defined reasoning mechanisms over structured knowledge. The reasoning processes implemented in this module are essential for answering user queries or generating specific outputs.

Explanations: Generates different types of explanations associated with specific queries related to answers or outputs of the system. It leverages outputs from the Symbolic Reasoning module (via the Query Answering module) and the Subsymbolic Services module.

Human in the Loop: In socio-technical environments, the system's effectiveness relies on adequately addressing user demands. This module aims to enhance system performance by incorporating iterative feedback from human users [24, 6, 35]. This feedback includes queries, responses, explanation requests, explanation ratings, utility-based classification of data sources, and argumentative exchanges, among other options.

Query Answering: Focuses on answering user queries by coordinating the execution of all other modules.

This is a general architecture that can be used for reasoning about different kinds of malicious behavior such as dissemination of fake news, hate speech and malware, detection of botnet operations, and prevention of cyber attacks including those targeting software products or blockchain transactions, among others. We briefly describe now specific instances of this framework used in the context of different malicious uses of social media.

4.1 Adversarial Deduplication

In real-world scenarios, a phenomenon related to (under) over-specification of information arises such as that of entity identification or deduplication, where the combination of an open-world assumption, conflicting information due to knowledge integration, and uncertain information makes it difficult to understand when different

[9] https://www.chistera.eu/projects/itrust

[10] A general-purpose software structure designed to facilitate the development of applications via instantiations or extensions.

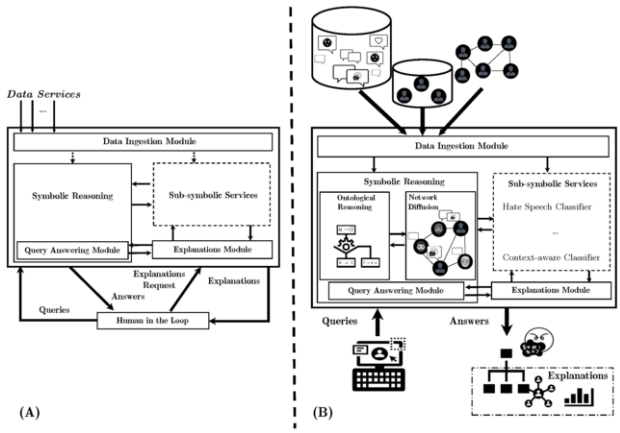

Figure 2: The HEIST (*Hybrid Explainable and Interpretable Socio-Technical systems*) application framework (A)[39], and its instantiation for detection of online hate speech (B)[29].

descriptions (profiles) of a user correspond to the same real-world entity. In [31, 32] we focus on this problem in settings where this multiplicity of entities is due to the fact that real-world actors do not want to be identified, such as malicious hacker forums and markets in which the participants are motivated to remain semi-anonymous; we coined the term *Adversarial Deduplication* towards this end. We studied this problem via examples that arise from real-world data on malicious hacker forums and markets arising from collaborations with a cyber threat intelligence company focusing on understanding this kind of behavior. In [32] we developed a set of experiments based on training ML classifiers that leverage text analysis to detect potential cases of duplicate entities. In [31] we propose a knowledge-based model that uses probabilistic existential rules (Probabilistic Datalog+/– [21]) to generate deduplication hypotheses under uncertainty. The main advantage with respect to existing deduplication tools (both based on statistical correlation and ML) is that our model operates under the open-world assumption, and thus is capable of modeling hypotheses over unknown objects (via value invention – nulls), which can later become known if new data becomes available.

4.2 Spreading of Fake News

Another malicious use of social media that is commonly observed is the spread of fake news. This social phenomenon has the potential to influence the opinions of millions of people who can be voters, consumers, or simply citizens going about their daily lives. In [29] we implemented and carried out an empirical evaluation of the HEIST framework described above for hybrid AI decision-support systems with the capability of leveraging the availability of ML modules, logical reasoning about unknown objects, and forecasts based on diffusion processes. We focus on the case of fake news dissemination on social platforms by three different kinds of users: non-malicious, malicious, and botnet members. In particular, we focus on three tasks: (i) determining who is responsible for posting a fake news article, (ii) detecting malicious users, and (iii) detecting which users belong to a botnet designed to disseminate fake news. Given the difficulty of obtaining adequate data with ground truth, we also developed a testbed that combines real-world fake news datasets with synthetically generated networks of users and fully detailed traces of their behavior throughout a series of time points. We designed our testbed to be customizable for different problem sizes and settings, and make its code publicly available to be used in similar evaluation efforts. We conducted a thorough experimental evaluation of three variants of our model and six environmental settings over the three tasks, to show the effects that the quality of knowledge engineering tasks, the quality of the underlying ML classifier used to detect fake news, and the specific environmental conditions have on smart policing efforts in social platforms.

4.3 Supervisor Agents

In [20] we investigated the design and implementation challenges faced in the deployment of a multi-agent system that operates in social network platforms to prevent or mitigate cyber attacks through the processing of streaming information. We instantiate the multi-agent system using the HEIST framework, which guides the implementation of hybrid socio-technical systems with a focus on explainability and discuss the main challenges in this process. We propose two possible approaches to building new knowledge dynamics operators: a cautious operator and a credulous operator, and evaluate the implications and challenges in each case.

The main goal of our model, a multi-agent system of Supervisor Agents, is to supervise social platforms, seeking to detect malicious content and activities and respond to avoid or mitigate their effect. We envision the use of such a system in examples as follows as described in [20]:

Medical content. A supervising system should be able to distinguish between a post with sexual content and a post that mentions sexual matters in a medical/health context. For example, it should prevent censorship of content related to breast cancer awareness—this would reduce false positives of sexual content on the social network. It could adjust alerts for dangerous/suspicious profiles against accounts that are whitelisted because they are known to disseminate alerts, educational content, awareness campaigns, etc. Currently, campaigns for breast cancer prevention cannot be freely shared as social networks censor any image of a breast, hindering the dissemination of proper self-examination and warning signs.

Parental control. Supervising systems can also be leveraged as tools that can be applied by users themselves in specific platforms to exert personalized control. Such a system could be conceived as an extension to be used "on top of" the social platforms, as is the case with Google's Family Link[11]. For instance, an application for mobile devices could, based on what is displayed on the screen, show alerts or—in the case in which the user is a minor—send notifications to guardians.

In order to tailor the HEIST framework for this purpose, we instantiated some of its modules as follows:

Data Ingestion. This component receives all the activity from the social platform, which includes all the events generated by all users. The stream activity is continuous and unbounded, so this module must deal with aspects related to stream processing such as windowing, load shedding, etc. [10]. As these tasks are completed, the module divides the data flow into windows, which are fed to the Symbolic Reasoning module.

Sub-symbolic Services. This module provides support in the form of basic services, such as user classification to predict certain behaviors in users [17], determining if posts contain hate speech, predicting the virality of posts, etc. This will allow for making more relevant security decisions, and deploying specific services depending on the context, such as image, audio, or video-based classifiers.

[11] https://families.google/familylink/

Symbolic Reasoning. This module takes input from the Data Ingestion module and is thus responsible for implementing the stream reasoning [11], maintaining and updating the agent knowledge base with the objective of detecting malicious behavior as events are processed through the sub-symbolic services. Rule-based approaches such as [36], or other formalisms based on computational logic, are good candidates for implementing such functionalities.

We are currently working in depth on the symbolic reasoning module, developing adequate logic-based reasoning engines. To quickly identify these threats and respond promptly and proactively, cybersecurity tools must be able to process the platform's data flow in as close to real-time as possible, which requires reasoning while avoiding bottlenecks and effectively handling activity peaks that are typical of these platforms.

5 Conclusions

In this paper, we have highlighted a series of proposals to integrate learning, knowledge, and reasoning in order to tackle problems associated with the malicious use of social media. Although most of these solutions are yet to be put in practice, preliminary evaluations show that these kinds of hybrid proposals are a promising alternative for the developing of socio-technical systems that can help to combat such threats minimizing risks to vulnerable populations, while promoting and protecting both individuals and collective rights.

Acknowledgments

This research has been partially supported by following projects: PID2022-139835NB-C21 funded by MCIN/AEI/10.13039/501100011033, PIE 20235AT010 and iTrust (PCI2022-135010-2).

References

[1] Coordinated inauthentic behavior, 2018. https://about.fb.com/news/tag/coordinated-inauthentic-behavior/.

[2] M. Ashida and M. Komachi. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of Sixth Workshop on Online Abuse and Harms (WOAH)*, July 2022.

[3] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proc. of 13th International Workshop on Semantic Evaluation*.

[4] S. Benesch. Countering dangerous speech: New ideas for genocide prevention. United States Holocaust Memorial Museum, 2014.

[5] H. Bonaldi, S. Dellantonio, S. S. Tekiroğlu, and M. Guerini. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *EMNLP*, Dec. 2022.

[6] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[7] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *ACL*, July 2019.

[8] Y.-L. Chung, M. Guerini, and R. Agerri. Multilingual counter narrative type classification. In *Proc. of the 8th Workshop on Argument Mining*, pages 125–132, Punta Cana, Dominican Republic, 2021. ACL.

[9] Y.-L. Chung, S. S. Tekiroğlu, and M. Guerini. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the ACL-IJCNLP 2021*, Aug. 2021.

[10] G. Cugola and A. Margara. Processing flows of information: From data stream to complex event processing. *ACM Computing Surveys*, 2012.

[11] E. Della Valle, S. Ceri, F. Van Harmelen, and D. Fensel. It's a streaming world! reasoning upon rapidly changing information. *IEEE Intelligent Systems*, 24(6):83–89, 2009.

[12] M. Fanton, H. Bonaldi, S. S. Tekiroğlu, and M. Guerini. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *ACK*, Aug. 2021.

[13] D. Furman, P. Torres, J. Rodríguez, D. Letzen, V. Martínez, and L. Alonso Alemany. Which argumentative aspects of hate speech in social media can be reliably identified? In *Proc.of 4th International Workshop on Designing Meaning Representations, IWCS 2023*, 2023.

[14] D. A. Furman, P. Torres, J. A. Rodríguez, D. Letzen, M. V. Martinez, and L. A. Alemany. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2942–2956, 2023.

[15] E. Gajewska, K. Budzynska, B. Konat, M. Koszowy, K. Kiljan, M. Uberna, and H. Zhang. Ethos and pathos in online group discussions: Corpora for polarisation issues in social media. *CoRR*, abs/2404.04889, 2024.

[16] F. R. Gallo, G. I. Simari, M. V. Martinez, M. A. Falappa, and N. A. Santos. Reasoning about sentiment and knowledge diffusion in social networks. *IEEE Internet Comput.*, 21(6):8–17, 2017.

[17] F. R. Gallo, G. I. Simari, M. V. Martinez, and M. A. Falappa. Predicting user reactions to twitter feed content based on personality type and social cues. *Future Generation Computer Systems*, 110:918–930, 2020.

[18] F. R. Gallo, G. I. Simari, M. V. Martinez, N. A. Santos, and M. A. Falappa. Local belief dynamics in network knowledge bases. *ACM Transactions on Computational Logic (TOCL)*, 23(1):1–36, 2021.

[19] F. R. Gallo, G. I. Simari, M. V. Martinez, N. A. Santos, and M. A. Falappa. Local belief dynamics in network knowledge bases. *ACM Trans. Comput. Log.*, 23(1):4:1–4:36, 2022.

[20] A. C. Garcia, C. A. D. Deagustini, J. C. Teze, M. V. Martinez, and G. I. Simari. A multi-agent system for addressing cybersecurity issues in social networks. In *Proc. of ENIGMA Workshop, KR 2023*, pages 43–54, 2023.

[21] G. Gottlob, T. Lukasiewicz, M. V. Martinez, and G. I. Simari. Query answering under probabilistic uncertainty in datalog+ / - ontologies. *Ann. Math. Artif. Intell.*, 69(1):37–72, 2013.

[22] R. Gupta, S. Desai, M. Goel, A. Bandhakavi, T. Chakraborty, and M. S. Akhtar. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proc. of 61st Annual Meeting of the Association for Computational Linguistics*, pages 5792–5809. ACL, 2023.

[23] M. Hinton and J. H. M. Wagemans. How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the GPT-3 AI text generator. *Argument Comput.*, 14(1):59–74, 2023.

[24] W. B. Knox and P. Stone. Augmenting reinforcement learning with human feedback. In *ICML 2011 Workshop on New Developments in Imitation Learning (July 2011)*, volume 855, page 3, 2011.

[25] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, July 2004.

[26] E. Mumford. The story of socio-technical design: Reflections on its successes, failures and potential. *Inf. Syst. J.*, 16:317–342, 2006.

[27] OSG. Social media and youth mental health: The US Surgeon General's advisory. Online (accessed 19-June-2024) – https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf, 2023.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, July 2002.

[29] J. Paredes, G. I. Simari, M. V. Martinez, and M. A. Falappa. Detecting malicious behavior in social platforms via hybrid knowledge- and data-driven systems. *Future Gener. Comput. Syst.*, 125:232–246, 2021.

[30] J. Paredes, J. C. Teze, M. V. Martinez, and G. I. Simari. The HEIC application framework for implementing xai-based socio-technical systems. *Online Soc. Networks Media*, 32:100239, 2022.

[31] J. N. Paredes, M. V. Martinez, G. I. Simari, and M. A. Falappa. Leveraging probabilistic existential rules for adversarial deduplication. In *Proceedings of PRUV@IJCAR 2018*. CEUR-WS, 2018.

[32] J. N. Paredes, G. I. Simari, M. V. Martinez, and M. A. Falappa. First steps towards data-driven adversarial deduplication. *Information*, 9(8): 189, 2018.

[33] J. Qian, A. Bethke, Y. Liu, E. M. Belding, and W. Y. Wang. A benchmark dataset for learning to intervene in online hate speech. *CoRR*, abs/1909.04251, 2019.

[34] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435.

[35] A. Rago, H. Li, and F. Toni. Interactive explanations by conflict resolution via argumentative exchanges. In *Proc. of the 20th International Conference on Principles of Knowledge Representation and Reasoning,*

KR 2023, pages 582–592, 2023.

[36] A. Ronca, M. Kaminski, B. C. Grau, and I. Horrocks. The delay and window size problems in rule-based stream reasoning. *Artificial Intelligence*, 306:103668, 2022.

[37] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. MIT Press, 2014.

[38] S. S. Tekiroğlu, Y.-L. Chung, and M. Guerini. Generating counter narratives against online hate speech: Data and strategies. In *Proc. of ACL 2020*, pages 1177–1190, 2020.

[39] J. C. L. Teze, J. Paredes, M. V. Martinez, and G. I. Simari. Engineering user-centered explanations to query answers in ontology-driven socio-technical systems. *Semantic Web*, pages 1–30 (*In Press*), 2024. doi: DOI:10.3233/SW-233297. URL https://content.iospress.com/articles/semantic-web/sw233297.

[40] L. Thorburn and A. Kruger. Optimizing language models for argumentative reasoning. In *Proc. of the 1st Workshop on Argumentation Machine Learning (COMMA 2022)*, 2022.

[41] M. E. Vallecillo Rodríguez, M. V. Cantero Romero, I. Cabrera De Castro, A. Montejo Ráez, and M. T. Martín Valdivia. CONAN-MT-SP: A Spanish corpus for counternarrative using GPT models. In *Proc. of LREC-COLING 2024*, pages 3677–3688, 2024.

[42] M. Vallecillo-Rodríguez, A. Montejo Ráez, and M. Martín-Valdivia. Automatic counter-narrative generation for hate speech in spanish. *Procesamiento del Lenguaje Natural*, 71, 2023.

[43] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. doi: 10.1126/science.aap9559. URL https://www.science.org/doi/10.1126/science.aap9559.

[44] J. H. M. Wagemans. Constructing a periodic table of arguments. In *Proceedings of 11th International Conference of the Ontario Society for the Study of Argumentation*, 2016.

[45] D. Walton, C. Reed, and F. Macagno. *Argumentation Schemes*. CUP, 2008.

[46] C. Ziems, B. He, S. Soni, and S. Kumar. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. 05 2020.