University of South Carolina
Scholar Commons

Publications

Artificial Intelligence Institute

Spring 4-21-2021

Defining and Detecting Toxicity on Social Media: Context and Knowledge are Key

Amit Sheth
University of South Carolina - Columbia, amit@sc.edu

Valerie Shalin
Wright State University

Ugur Kursuncu
University of South Carolina - Columbia, ugurkursuncu@gmail.com

Follow this and additional works at: https://scholarcommons.sc.edu/aii_fac_pub

Part of the Anthropology Commons, Communication Commons, Computer Engineering Commons, Digital Humanities Commons, Electrical and Computer Engineering Commons, Library and Information Science Commons, Other Languages, Societies, and Cultures Commons, Other Social and Behavioral Sciences Commons, Psychology Commons, Religion Commons, and the Sociology Commons

Publication Info
Preprint version *Neurocomputing*, Spring 2021.
© The Authors, 2021.

This Article is brought to you by the Artificial Intelligence Institute at Scholar Commons. It has been accepted for inclusion in Publications by an authorized administrator of Scholar Commons. For more information, please contact dillarda@mailbox.sc.edu.

Defining and Detecting Toxicity on Social Media: Context and Knowledge are key

Amit Sheth[1], Valerie L. Shalin[2], Ugur Kursuncu[1]

[1] AI Institute, University of South Carolina

[2] Department of Psychology, Wright State University

Abstract

As the role of online platforms has become increasingly prominent for communication, toxic behaviors, such as cyberbullying and harassment, have been rampant in the last decade. On the other hand, online toxicity is multi-dimensional and sensitive in nature, which makes its detection challenging. As the impact of exposure to online toxicity can lead to serious implications for individuals and communities, reliable models and algorithms are required for detecting and understanding such communications. In this paper We define toxicity to provide a foundation drawing social theories. Then, we provide an approach that identifies multiple dimensions of toxicity and incorporates explicit knowledge in a statistical learning algorithm to resolve ambiguity across such dimensions.

1. Introduction

Online social media platforms are arguably the most culturally significant technological innovations of the 21st century. The numerous benefits include the wide distribution of content crossing geographic boundaries, and enabling interaction and exchanges that are nearly free of physical constraints except infrastructure [1]. Communities have emerged around every conceivable special interest from science to travel, from politics to child rearing. The easy spread of data, information, and knowledge were expected to benefit the ability to foster informed decision making, cultural exchanges and the coordination of activities online and in the physical world. Unfortunately, social media has also significantly enhanced the reach and scale of harmful content including disinformation, conspiracies, extremism, harassment, violence, and other forms of socially toxic material [2, 3]. While social media platforms attempt to counter and overcome such harmful content and behavior, their efforts are largely ineffective and as such themselves have the potential for unintended adverse impact. The effort and effectiveness of moderation is potentially subject to the companies' economic interest, political and regulatory considerations, or due to the lack of effective tools and sufficient investment in the effort. Human content moderation has resulted in relatively unsatisfactory outcomes [4]. The political and public health climate of 2020 encouraged society to adopt technological and specifically AI-based solutions with limited understanding. A prominent reason is the lack of understanding the challenging nature of toxicity, which fundamentally requires context outside of the explicit content. The detection of toxicity demands an interdisciplinary perspective with empirical approaches. Consistent with our people content network framework for the characterization of social media exchange [5, 6], we assert the more general role of context, and in particular cultural context, in the interpretation of content. This paper identifies three issues:

- identify the psychological and social dimensions of the problem
- identify the limitations of contemporary computational approaches, and
- outline an advanced technical approach founded on knowledge-driven context based analysis.

2. A PsychoSocial problem Meets Computation

Our view of toxic content extends beyond the currently used classification which focused on “threats, obscenity, insults, and identity-based hate” [1]. We also include harassment and socially disruptive persuasion, such as misinformation and radicalization. While the cultural foundations of toxicity are readily

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview

apparent in misinformation and radicalization, we contend that culture provides essential context to the determination of any toxic content. Figure 1 guides this section, starting with conventional content analysis and expanding the psychological, social and cultural scope of the required analysis.

2.1. Content Analysis

By far the most common approach to toxicity detection focuses on the content of exchanges. Offensive key words, often so called "coarse language" are easy to tabulate in a lexicon. More sophisticated analyses employ lexicons specific to intelligence, appearance, race, sexual preference etc. [7]. Keyword based content analysis encounters a number of challenges. An evolving culture conveys an insulting connotation to otherwise apparently banal language, e.g., basic, cancel, Karen, shade, snowflake, and thirsty. This not only requires constant maintenance of the lexicon, but context to disambiguate the slang usage from general usage [8].

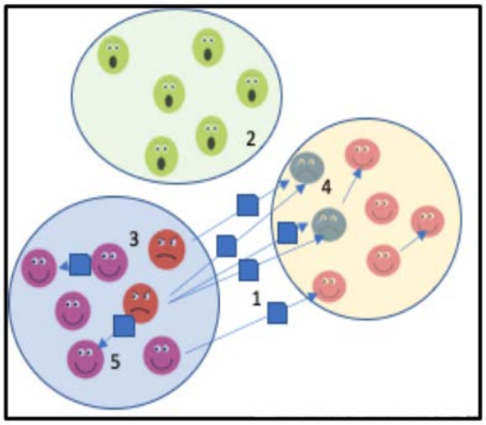

Figure 1: Conventional toxicity analysis examines the content exchanged between individuals in a community. Often external observers impose their own culturally biased decision rules. Detecting toxic sources expands analysis, but still fails to acknowledge the reaction of the target, which is likely tempered by common group membership.

A second problem is that content analysis based on isolated lexical items does not necessarily confer toxicity. For example, North American teenagers readily employ language among themselves that adults would consider offensive. More worrisome, the word "jihad" may be readily interpreted as a radical content by a Westerner, but a more culturally sensitive analysis reveals that this terms could also have been used in a benign religious text [9]. The scope of toxic topics, general knowledge and cultural foundations are virtually unbounded. "Dressing like your grandmother" directed to a teen is laden with cultural deprecation founded on both ageism and consumerism, but contains no single offensive word in isolation. Irony, humor, and teasing between friends precludes simple net sentiment analysis.

Toxic content is multimodal, often exploiting images and video that are not well monitored or understandable using contemporary technology [10]. Facial recognition algorithms fail miserably for dark skinned females. More generally, the benchmark image databases heavily favor western culture, at the expense of eastern cultures (China is represented by only 1% of images in the Imagenet[2]). Apart from these obvious multi-modal processing challenges, text and image content must be aligned in a common framework at the appropriate level of abstraction. Finally, we cannot assume that text provides useful image processing guidance. A cheery "Have a Nice Day" can easily pair with an embarrassing photo of the recipient.

2.2. Culturally bound decision criteria

Even a simple content analysis requires a decision criterion. Framing toxicity as a standard signal detection problem acknowledges the potential for two overlapping distributions of potentially toxic content instances, one over relatively low toxicity values and another over higher values. The decision criterion is vulnerable to cultural considerations. All-too-common annotator disagreement is resolved by the vote of a small sample of annotators, while the foundations of disagreement remain unstudied. Annotators may hold unconscious stereotypes, for example associating religion with radicalization. Personal experience and cultural differences create variable interpretations of label semantics; disagreement over the nature of misinformation is the source of great concern in moderating internet content.

We have already noted the preponderance of coarse language in teenagers [11, 12, 13]. The population of non-toxic content is much larger than the pop-

[2] https://devopedia.org/imagenet

ulation of toxic content [10, 14, 15]. Class imbalance does not work well with contemporary machine learning algorithms. As we lower the toxicity decision rule we admit more false positives, potentially resulting in orders of magnitude more false positives than true positives. This creates both adverse impact to the falsely accused along with a practical problem of follow up [9].

Finally, corpus assembly itself conveys cultural bias. The classification algorithm for one population does not generalize to different populations, creating a validity problem. For example, the relevant data set for detecting toxic exchange among adolescents needs to be different from the data set for defining such exchange among adults [10] due to different cultural practices.

2.3. Identifying the toxic source

The source's Intent to hurt or harm is the defining feature of bullying. Harm may employ the disclosure of sensitive facts, denigrate, be grossly offensive, be indecent or obscene, be threatening, make false allegations, deceive, spam, spread misinformation, mimic interest, clone profile or personal invade space. The motivation for detecting the toxic source is mitigation, but the false alarm risk is real. Moreover, one instance is unlikely to constitute sufficient evidence. Evaluation of the potentially toxic source requires a corpus of the candidate's content, raises a challenge of corpus scope and introduces an aggregation of evidence.

2.4. Identifying the target

Experienced harm is distinct from intent to harm, from the recipient's perspective resulting in discrimination, deception, fraud, diseducation, loss of money, offense, loss of reputation, manipulation, embarrassment, distraction, loss of time, fraud. Moreover, harassment, by definition refers to the special case of bullying with respect to a protected class [16]. The nefarious source takes advantage of the specific features of target vulnerability such as age, occupation and public stature. Contemporary victims of bullying include Parkland High School

Students[3]. The motivation for detecting the target is protection. Both bullying and harassment are associated with cases of adolescent suicide [17, 18, 19]. Age matters in the assessment of target experience; adolescent brains are still developing an ability to process social feedback making adolescents particularly vulnerable to both negative feedback [20], and radicalization efforts [21].

2.5. Participants' Relationship and Group membership

Friendship, power differentials and social network membership provide essential context. Insults are common between adolescent friends. Participants with the same racial background readily exchange otherwise offensive racial epithets. Social network structure has at least two consequences founded on the distinction between in-group and out-group membership and the target's position within these groups. First, multiple negative messages from different participants in the in-group targeted to an out-group recipient are as potentially toxic as the same number of messages from a single source. Social network membership is therefore an important feature in the detection of toxicity. Second, the promise of group membership and the threat of exclusion is a known factor in the radicalization effort [22], of particular appeal to adolescent recruits.

As surrounding benign conversation mitigates the single potentially toxic comment, exchange history informs the determination of toxicity. Hence, exchange history surrounding the potentially toxic comment must be present in the corpus. Crawling on sender and recipient identifiers is too limited. Victims are often targeted with mention tags in an exchange between a sender and what might be charitably called bystanders. These concerns illustrate that the scope of an annotation item is problematic. A single episode may look quite different in the context of other exchanges, suggesting that the potentially benign or toxic instance should be annotated with respect to its broader historical and network context. This suggests currently unspecified requirements for sys-

[3] https://www.huffpost.com/entry/parkland-students-combat-cyberbullying$_{n5}aeeee86e4b033e5c3f03126$

tematically defining and scoping the annotation task. Expanded context also raises the problem of conflicting indicators, the assessment of stale content, and emphasizes the need for confidence estimates.

3. Technical Challenges to Automated Detection of Toxic Language

As described above, the toxicity detection problem is not a purely computer science or AI problem. Toxicity detection is an interdisciplinary problem founded on theory, empirical models and knowledge to guide classification [23, 10]. But to identify toxicity, it is necessary to understand the broader context beyond situation and domain specific content analysis, with reference to applicable human values, social norms and culture, at an individual, group and community levels. In contrast, conventional approaches for identification of toxic exchange have been treated as a content processing problem [24, 25]. The state-of-the-art algorithms used in modeling toxic content are mostly autoregressive models (e.g., BERT, GPT-2,3), which are designed to predict the next token given previous tokens from the dataset as input. As these models have been trained using data collected from the web, corpus bias and incidentally confounded features result in models that can cause intentional or unintentional harms to individuals or society[4] [26, 27, 3]. Recent studies [28, 29, 30] suggested that these state-of-the-art algorithms are prone to generating racist or sexist schemes. While these models can be retrained using transfer learning for the problem in hand by fine-tuning to update the model parameters, significant bias will still carry over, which might potentially cause harm. Recent studies demonstrated that these fine-tuned models can particularly be dangerous in highly sensitive areas, such as online toxicity as well as health [31, 32, 33]. For instance, Google's Perspective designed for toxicity detection received criticism for biased scoring of content based on gender, sexual orientation, religion or disability. Further, this model almost always assigned a high toxicity score

[4] https://thenextweb.com/neural/2021/01/19/gpt-3-has-consistent-and-creative-anti-muslim-bias-study-finds/

if the content included insults or profanity, regardless of the intent or tone of the author [34]. Hence, policymakers and practitioners assert serious usability and safety concerns that constrain adoption of these technologies that are not well-understood in terms of their impact on individuals and society [35].

In this section, we discuss the technical consequences of our expanded approach to toxicity detection in three subsections: the need for empirical models, the need for a curated corpus, and the need for external knowledge.

3.1. Need for Empirical Models

Computational modeling of human behavior often requires domain expertise to inform the classes and subclasses of toxicity. On the other hand, such domain expertise is scarce; hence, we need to provide a conceptual knowledge model in a structured or semi-structured format that is readable by machines. Three critical issues are the resolution of context, ambiguity and mitigating unfairness. Researchers often resort to the post-level approach for building datasets and design algorithms to detect toxicity between two individuals focusing on recognition of explicit language of insult [10]. Specifically on social media, posts are often short with inadequate information for context which represent substantial ambiguity. For example, a playful exchange between good friends with sarcastic content could be falsely flagged as harassment, or a religious reference to "jihad" could falsely flag a pious worshipper as an extremist. Below we consider the use cases for cursing, extremism and harassment to demonstrate the need for empirical models to guide analysis.

Cursing. The intention of the parties of a conversation along with social context, determines the meaning of their language. In [36], we studied the communications on Twitter concerning the use of cursing and its relations with intention and emotions. While we found around 8% of conversations contain profanity and curse words, the intention of users may not necessarily be toxic. We explored the role of emotions in identifying intention, as cursing may be associated with positive emotions as well as negative emotions, and these emotions may indicate the real intention. We identified three contextual variables that determine the social context as to *when, where* and *how* the cursing occurs. For example, we found that people curse more when they wake up, in relaxed virtual environments.

Extremism. In [9] we started with the notion in political science that radicalization is a process employed by extremist groups, with systematic changes in persuasive content over time. As the type of extremism we study was Islamist extremism, appropriate domain expertise is critical to distinguish the true extremist from non-extremist communication. Guided by an empirical model developed by a political scientist, we supported three dimensions to model this content: religion, ideology and hate. Ambiguity is a significant challenge as diagnostic terms in predicting extremism often have different meanings. For instance, the meaning of the keyword "jihad" in religion is referred to as a self-spiritual struggle, while it indicates intent to harm other individuals in the Islamist extremist ideology. As the same term has two different meanings for extremist and non-extremist content, it needs to be represented differently in a computational model for resolving such ambiguity. Hence, a multi-dimensional and contextual modeling of this content incorporating knowledge (in this example, the religious knowledge that help distinguish the two meanings) allows us to address ambiguity, reducing false alarm and mitigating unfairness. Considering a potential deployment of a socially responsible model with improved fairness would mitigate adverse impact on nearly 2 billion Muslims.

Harassment. Many early researchers defined harassment as a binary classification - a social media post (e.g., a tweet) is either harassing or not [37, 38, 39, 40]. As the context is crucial in capturing harassment, it will change based on the linguistic meaning, interpretation, and distribution. In [7] we offered more dimensions of harassment including; (i) sexual, (ii) racial, (iii) appearance-related, (iv) intellectual, and (v) political content, and created a type-aware lexicon and annotated dataset [41]. Then we employed a multi-class classification algorithm based on these five dimensions. While coarse lexical items signal some of these, ambiguous common language (fat, dumb) and idioms are also relevant. A multi-class approach is required because perpetrators can exploit more than one subclass in targeting a victim. In the absence of a multi-class model, the victim's experience over time will not surface.

3.2. Need for Curated Corpora

The analysis is only as good as the corpus. Researchers often resort to the post-level approach for building datasets and designing algorithms to detect toxicity between two individuals focusing on recognition of explicit language of insult [10]. Keyword based crawls ignore message context, and create ambiguity, e.g., a playful exchange between good friends with sarcastic content could be falsely flagged as harassment, or a religious reference to "jihad" could falsely flag a pious worshipper.

Extremism. For our extremism project, we relied upon a curated corpus [42] consisting of 538 verified extremist users, established by Twitter and the Lucky Troll club [43]. We balanced this with a corpus of 538 non-extremist users from an annotated Muslim religious dataset [44]. We make two points with this example. First, the set of positive cases reflected professional judgment. Second, the applicability of the resulting classification model depends on the quality of the distractor corpus. Here we were particularly concerned with adverse impact and therefore employed a distractor corpus that posed significant false alarm opportunity. Nevertheless, this balanced corpus does not reflect the class imbalance in the uncurated data. Even very high precision results can produce a large number of false alarms in an unbalanced corpus [9].

Harassment. We curated our own corpus for our high school harassment that addresses a number of corpus considerations [10], under an IRB approved protocol requiring privacy protections through anonymization. First, because the culture of the U.S. high school population is quite different from the general U.S. culture at large, we assured the identity of the participants. Starting with a seed set of known high school student names published in the newspaper as scholarship winners, we searched Twitter for unique matches to users with appropriate location indicators in their metadata. To grow the set, we searched on their Twitter contacts, and then pruned the resulting list of candidates by requiring contacts with other members of the candidate list. Second, we make no assumptions regarding the nature of toxic content in assembling this corpus. Third, as we were concerned with capturing the full context for the individual post, we retrieved the history of exchanges between members, and the multimodal content of these exchanges including emoji and images which may also contain toxic content [10]. The diversity of modality enriches the interactions between humans and computers. Specifically, users create the context of their conversations using these modalities making implicit relationships in between. As a result of our corpus assembly process, we can recover network structure [45] suitable for insider-outsider analysis. Finally, with the caveat of access restricted to public accounts, our corpus approximates a realistic class balance of benign and toxic content.

3.3. Need for External Knowledge

We advocate the use of relevant types of knowledge in a variety of forms, such as ground truth corpora and knowledge graphs (KGs). This assures attention to the different dimensions to understand subtle nuances in semantics/meaning of toxic behavior. External knowledge constitutes a source of “ground truth” for evaluating message content. As we argue that toxic behavior is multidimensional leading to ambiguity and false alarms, we employ a multi-level and multi-dimensional approach that helps capture differences between various cultural and societal understandings of toxicity resolving ambiguity. A framework for this broader context for interpretation and evaluation is offered in Purohit, et. al., [46], which identified three major dimensions of knowledge necessary to design humanity-inspired AI systems: personalization, social context and intention. Differentiating the users and their content requires different levels of granularity in organization of features. Our previous Person, Content, and Network (PCN) distinction [5, 6, 1] functions at a higher (superficial) level, whereas the contextual dimensions of content (e.g., religion, ideology and violence) functions at the lower level [9], capturing the deep semantics of toxicity. Further, when incorporated into a classification algorithm, external knowledge enables opportunities to provide an explanation generally missing from contemporary deep learning approaches [47, 48, 49].

Here, we expand citepurohit2020knowledge to scope the relevant knowledge which described three dimensions: values, norms and the domain. Each dimension is pegged by individual specificity and collective generality, and the perspective required to interpret the behavior of an individual is represented by the combination of all three dimensions. While citepurohit2020knowledge considered other actors as part of the environment, here, we consider them more explicitly. The concept of *personal semantics* for the target of toxicity covers much of what citepurohit2020knowledge intended in their analysis. Personal semantics includes knowledge about the targets' language of insult, verbal abuse and offensive language, involving sensitive topics specific to the individual and their social network. From the sources' perspective, we require knowledge corresponding to their *intention*, particularly associated with indicators of power, truth, and trust [50]. Finally, the *history of interaction* such as duration and toxicity frequency between source and target requires knowledge about the structure of nominal conversation such as indicators of topic change and common ground that determine familiarity [51]. The target's *emotional response* corresponds to the toxicity-specific emotion evoked in a recipient after reading messages, informed by conversation history and network membership. This requires a more sophisticated classification scheme beyond binary toxicity, referring to: the causes of experienced harm, embarrassment, loss of reputation, etc. as well as possible clinically relevant consequences such as depression and suicide [2]. While these sources of knowledge are typically not made explicit in toxicity analysis, the failure to make them explicit or acknowledge features corresponding to these contributes to disagreement among annotators and ultimately poor, and biased classification.

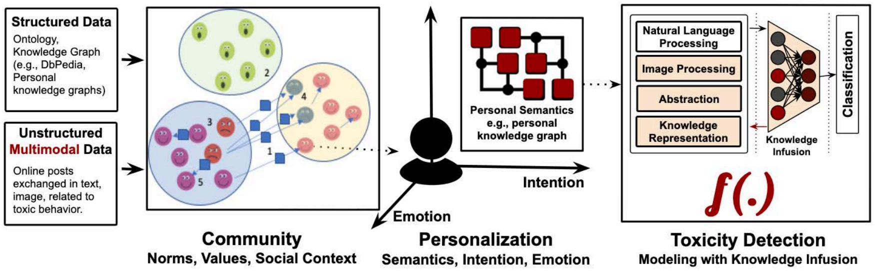
Structured Data
Ontology, Knowledge Graph (e.g., DbPedia, Personal knowledge graphs)
Unstructured Multimodal Data
Online posts exchanged in text, image, related to toxic behavior.
1
2
3
4
5
Community
Norms, Values, Social Context
Personal Semantics e.g., personal knowledge graph
Intention
Emotion
Personalization
Semantics, Intention, Emotion
Natural Language Processing
Image Processing
Abstraction
Knowledge Representation
Knowledge Infusion
Classification
f(.)
Toxicity Detection
Modeling with Knowledge Infusion

Figure 2: Individuals are surrounded by the sources of data and knowledge required for computational analysis, personalizing the analysis by incorporating personal semantics, intention and emotion will help better distinguish toxic behaviors from non-toxic. Further, infusing external knowledge will resolve ambiguity better contextualizing multimodal data and provide a source of explanation.

4. A Knowledge-enhanced Socio-technical Approach to Toxicity Detection

Toxicity detection takes the form of two problems: detection of the toxic source and identification of the vulnerable victim. In either case, we require more sophisticated Natural Language Processing (NLP) and Machine Learning (ML) methods to both detect and use the features indicative of toxicity. Specifically, the meaning of content changes based on the belief system of the source and target; thus, they need to be computationally represented separately. These personalized belief systems are critical for understanding how toxic behavior is interpreted differently by different individuals [52]. These inter-related concepts and beliefs also evolve over time upon exposure to new information [53]. The question here is how one can computationally model the evolution of complex social exchange. We advocate a Knowledge-infused Learning (K-iL) framework where the model learns to recognize patterns of different meanings of toxic concepts from different perspectives to reduce ambiguity. However, the knowledge sources are not necessarily at the same level of granularity and abstraction. Accordingly, we categorized knowledge infusion as shallow, semi-deep and deep infusion [54, 9] described below, to resolve the impedance mismatch due to different representational forms and abstractions. Infusing knowledge is particularly important for overcoming the inescapable limitations and biases of data-driven processing [55].

We propose a framework that will account for Purohit et al.'s dimensions to generate richer representations including personal semantics, intention, emotion, history of interaction and social context. This collection of information will require a dynamic hybrid design that will cultivate models for different modalities of data and knowledge representation. As behavioral models are dynamic and evolve, this framework should also allow for change. Lastly, the fusion of representations generated from these models that represent multiple dimensions also poses a new computational challenge. Further, validation of such an approach is also challenging and likely requires some form of experimentally controlled data collection to support supervised learning. This framework will address multiple levels of data, such as content, individual and community, ensuring that the individual level details are changing as interacted with their network. Communities are formed around various topics of interest through network interactions, where the shared content displays an intent attached with emotions. Hence, the individuals in toxic interactions (e.g., aggressor, victim) show different characteristics, and it is critical to bring to bear different dimensions, such as content, individual and network, for reliable analysis. As learning concepts and grasping causal relations go beyond the data available, conceptual and probabilistic models can perform inference over hierarchies of structured representations [56].

Among the Purohit et al. dimensions, personal semantics, interactions and social context can be represented using both conceptual (e.g., knowledge graphs) and probabilistic models (e.g., language, image). External knowledge can be represented in structured (e.g., knowledge graph) and semi-structured forms (e.g., JSON) to inform computation. While knowledge can be acquired from data through various methods, dependence over data significantly limits search space and extraction of the complete knowledge that is required to represent the complex nature of toxicity [57]. Explicit structural relations in a knowledge graph constitute context and capture the intrinsic characteristics of this problem, which can be incorporated into a statistical learning algorithm (e.g., neural networks) to enhance the contextual latent space. This incorporation will adjust emphasis on sparse-but-essential and irrelevant-but-frequent terms and concepts, boosting recall without reducing precision. While probabilistic models (e.g., BERT, GPT-3, ResNet, Inception) have advanced in recent years, generating knowledge representations from knowledge graphs or similar structured forms of knowledge remains an open area for advancement. However, a knowledge graph can be generated as embedding vectors including structural information of the graph, such as relationships. Existing methods, such as TRANS-E [58], TRANS-H [59], and HOLE [60], can be utilized to generate embeddings from a knowledge graph. The generated knowledge representation can, then, be infused within a probabilistic model.

In a learning architecture, represented knowledge can be infused through an attention mechanism and knowledge-based constraints or dependency relations between words in a sentence [61]. Deep infusion of knowledge is still an open area of research, and we described our approach in [54]. Deep infusion of knowledge combines the representation of structural knowledge graph content with a latent representation of data, quantifying the information loss and identifying the level of abstraction. The infusion of knowledge can take place after each epoch optimizing the loss function. In this architecture, for deep infusion, related functions add an additional layer which takes the latent vectors of the previous layers, and the knowledge embedding, merging them to output a knowledge infused representation. In this framework, as we utilize multiple dimensions to represent toxic behavior, appropriate infusion of knowledge will form connections within the data resulting in better contextualized representation. As our prior work suggests, infusion of knowledge mitigates unfair outcomes by reducing false positives which would lead to adverse societal implications [9, 54, 2, 62].

5. Conclusion

In this paper, we identified the multiple influences on the detection of toxic exchange beyond conventional content analysis. Our goal was to provide a framework that identifies and utilizes multiple dimensions of toxicity and incorporates explicit knowledge in a statistical learning algorithm to resolve ambiguity across such dimensions. Specifically, we highlighted the significance of multi-level analysis of data, namely, content, individual and community, and the features necessary to determine toxicity. Knowledge representation and its infusion in a learning algorithm is an emergent solution for toxicity detection and related sets of similar problems. For toxicity detection we provided a framework founded on behavioral and social theory.

Acknowledgement

This work is funded in part by National Science Foundation Award 1761931 Spokes: MEDIUM: MIDWEST: Collaborative: Community-Driven Data Engineering for Substance Abuse Prevention in the Rural Midwest. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] U. Kursuncu, M. Gaur, U. Lokala, K. Thirunarayan, A. Sheth, I. B. Arpinar, Predictive analysis on twitter: Techniques and applications, in: Emerging research challenges and opportunities in computational social network analysis and mining, Springer, 2019, pp. 67–104.

[2] U. Kursuncu, H. Purohit, N. Agarwal, A. Sheth, When the bad is good and the good is bad: Understanding cyber social health through online behavioral change, IEEE Internet Computing 25 (01) (2021) 6–11.

[3] U. Kursuncu, Y. Mejova, J. Blackburn, A. Sheth, Cyber social threats 2020 workshop meta-report: Covid-19, challenges, methodological and ethical considerations, Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media.

[4] T. Gillespie, Content moderation, ai, and the question of scale, Big Data & Society 7 (2) (2020) 2053951720943234.

[5] H. Purohit, Y. Ruan, A. Joshi, S. Parthasarathy, A. Sheth, Understanding user-community engagement by multi-faceted features: A case study on twitter, in: WWW 2011 Workshop on Social Media Engagement (SoME), 2011.

[6] U. Kursuncu, M. Gaur, U. Lokala, A. Illendula, K. Thirunarayan, R. Daniulaityte, A. Sheth, I. B. Arpinar, What's ur type? contextualized classification of user types in marijuana-related communications using compositional multiview embedding, in: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, 2018, pp. 474–479.

[7] M. Rezvan, S. Shekarpour, F. Alshargi, K. Thirunarayan, V. L. Shalin, A. Sheth, Analyzing and learning the language for different types of harassment, Plos one 15 (3) (2020) e0227330.

[8] A. Sheth, P. Kapanipathi, Semantic filtering for social data, IEEE Internet Computing 20 (4) (2016) 74–78.

[9] U. Kursuncu, M. Gaur, C. Castillo, A. Alambo, K. Thirunarayan, V. Shalin, D. Achilov, I. B. Arpinar, A. Sheth, Modeling islamist extremist communications on social media using contextual dimensions: religion, ideology, and hate, Proceedings of the ACM on Human-Computer Interaction 3 (CSCW) (2019) 1–22.

[10] T. Wijesiriwardene, H. Inan, U. Kursuncu, M. Gaur, V. L. Shalin, K. Thirunarayan, A. Sheth, I. B. Arpinar, Alone: A dataset for toxic behavior among adolescents on twitter, in: International Conference on Social Informatics, Springer, 2020, pp. 427–439.

[11] T. Jay, Cursing in America: A Psycholinguistic Study of Dirty Language in the Courts, in the Movies, in the Schoolyards, and on the Streets, John Benjamins Publishing, 1992.

[12] T. Jay, The utility and ubiquity of taboo words, Perspectives on psychological science 4 (2) (2009) 153–161.

[13] M. R. Mehl, J. W. Pennebaker, The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations., Journal of personality and social psychology 84 (4) (2003) 857.

[14] Z. Waseem, J. Thorne, J. Bingel, Bridging the gaps: Multi task learning for domain transfer of hate speech detection, in: Online harassment, Springer, 2018, pp. 29–55.

[15] J. Golbeck, Z. Ashktorab, R. O. Banjo, A. Berlinger, S. Bhagwan, C. Buntain, P. Cheakalos, A. A. Geller, R. K. Gnanasekaran, R. R. Gunasekaran, et al., A large labeled corpus for online harassment research, in: Proceedings of the 2017 ACM on web science conference, 2017, pp. 229–233.

[16] S. Einarsen, H. Hoel, D. Zapf, C. Cooper, Bullying and harassment in the workplace: Developments in theory, research, and practice, CRC press, 2010.

[17] S. Hinduja, J. W. Patchin, Bullying, cyberbullying, and suicide, Archives of suicide research 14 (3) (2010) 206–221.

[18] W. B. Roberts, D. H. Coursol, Strategies for intervention with childhood and adolescent victims of bullying, teasing, and intimidation in school settings, Elementary School Guidance & Counseling 30 (3) (1996) 204–212.

[19] G. D. Cooper, P. T. Clements, K. E. Holt, Examining childhood bullying and adolescent suicide: Implications for school nurses, The Journal of School Nursing 28 (4) (2012) 275–283.

[20] E. A. Crone, E. A. Konijn, Media use and brain development during adolescence, Nature communications 9 (1) (2018) 1–10.

[21] W. Pedersen, V. Vestel, A. Bakken, At risk for radicalization and jihadism? a population-based study of norwegian adolescents, Cooperation and conflict 53 (1) (2018) 61–83.

[22] S. Ozer, M. Obaidi, S. Pfattheicher, Group membership and radicalization: A cross-national investigation of collective self-esteem underlying extremism, Group Processes & Intergroup Relations 23 (8) (2020) 1230–1248.

[23] S. Henry, School violence beyond columbine: A complex problem in need of an interdisciplinary analysis, American Behavioral Scientist 52 (9) (2009) 1246–1265.

[24] D. Noever, Machine learning suites for online toxicity detection, arXiv preprint arXiv:1810.01869.

[25] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, I. Androutsopoulos, Toxicity detection: Does context really matter?, arXiv preprint arXiv:2006.00998.

[26] K. McGuffie, A. Newhouse, The radicalization risks of gpt-3 and advanced neural language models, arXiv preprint arXiv:2009.06807.

[27] A. Olteanu, C. Castillo, F. Diaz, E. Kıcıman, Social data: Biases, methodological pitfalls, and ethical boundaries, Frontiers in Big Data 2 (2019) 13.

[28] S. Gehman, S. Gururangan, M. Sap, Y. Choi, N. A. Smith, Realtoxicityprompts: Evaluating neural toxic degeneration in language models, arXiv preprint arXiv:2009.11462.

[29] S. Groenwold, L. Ou, A. Parekh, S. Honnavalli, S. Levy, D. Mirza, W. Y. Wang, Dats wassup!!: Investigating african-american vernacular english in transformer-based text generation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 5877–5883.

[30] E. Wallace, S. Feng, N. Kandpal, M. Gardner, S. Singh, Universal adversarial triggers for attacking and analyzing nlp, arXiv preprint arXiv:1908.07125.

[31] H. Zhang, A. X. Lu, M. Abdalla, M. McDermott, M. Ghassemi, Hurtful words: quantifying biases in clinical contextual word embeddings, in: proceedings of the ACM Conference on Health, Inference, and Learning, 2020, pp. 110–120.

[32] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proceedings of the National Academy of Sciences 115 (16) (2018) E3635–E3644.

[33] I. Y. Chen, P. Szolovits, M. Ghassemi, Can ai help reduce disparities in general medical and mental health care?, AMA journal of ethics 21 (2) (2019) 167–179.

[34] L. Hanu, J. Thewlis, S. Haco, How ai is learning to identify toxic online content, Scientific American.

[35] E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature medicine 25 (1) (2019) 44–56.

[36] W. Wang, L. Chen, K. Thirunarayan, A. P. Sheth, Cursing in english on twitter, in: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, 2014, pp. 415–425.

[37] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, L. Edwards, Detection of harassment on web 2.0, Proceedings of the Content Analysis in the WEB 2 (2009) 1–7.

[38] G. Kennedy, A. McCollough, E. Dixon, A. Bastidas, J. Ryan, C. Loo, S. Sahay, Technology solutions to combat online harassment, in: Proceedings of the first workshop on abusive language online, 2017, pp. 73–77.

[39] A. Bastidas, E. Dixon, C. Loo, J. Ryan, Harassment detection: a benchmark on the# hackharassment dataset, arXiv preprint arXiv:1609.02809.

[40] M. Bugueño, M. Mendoza, Learning to detect online harassment on twitter with the transformer., Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

[41] M. Rezvan, S. Shekarpour, L. Balasuriya, K. Thirunarayan, V. L. Shalin, A. Sheth, A quality type-aware annotated corpus and lexicon for harassment research, in: Proceedings of the 10th ACM Conference on Web Science, 2018, pp. 33–36.

[42] M. Fernandez, M. Asif, H. Alani, Understanding the roots of radicalisation on twitter, in: Proceedings of the 10th acm conference on web science, 2018, pp. 1–10.

[43] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, A. Galstyan, Predicting online extremism, content adopters, and interaction reciprocity, in: International conference on social informatics, Springer, 2016, pp. 22–39.

[44] L. Chen, I. Weber, A. Okulicz-Kozaryn, Us religious landscape on twitter, in: International Conference on Social Informatics, Springer, 2014, pp. 544–560.

[45] S. Bhatt, S. Padhee, A. Sheth, K. Chen, V. Shalin, D. Doran, B. Minnery, Knowledge graph enhanced community detection and characterization, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 51–59.

[46] H. Purohit, V. L. Shalin, A. P. Sheth, Knowledge graphs to empower humanity-inspired ai systems, IEEE Internet Computing 24 (4) (2020) 48–54.

[47] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, T.-S. Chua, Explainable reasoning over knowledge graphs for recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 5329–5336.

[48] F. Lecue, On the role of knowledge graphs in explainable ai, Semantic Web 11 (1) (2020) 41–51.

[49] M. Gaur, K. Faldu, A. Sheth, Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable?, IEEE Internet Computing 25 (1) (2021) 51–59.

[50] R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust, Academy of management review 20 (3) (1995) 709–734.

[51] I. Hutchby, Conversation analysis, The Wiley-Blackwell Encyclopedia of Social Theory (2017) 1–9.

[52] N. E. Friedkin, A. V. Proskurnikov, R. Tempo, S. E. Parsegov, Network science on belief system dynamics under logic constraints, Science 354 (6310) (2016) 321–326.

[53] J. L. Usó-Doménech, J. Nescolarde-Selva, What are belief systems?, Foundations of Science 21 (1) (2016) 147–152.

[54] U. Kursuncu, M. Gaur, A. Sheth, Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning, Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice (AAAI-MAKE 2020).

[55] A. P. Sheth, K. Thirunarayan, The inescapable duality of data and knowledge, 2021.

[56] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, How to grow a mind: Statistics, structure, and abstraction, science 331 (6022) (2011) 1279–1285.

[57] L. G. Valiant, Robust logics, Artificial Intelligence 117 (2) (2000) 231–253.

[58] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Neural Information Processing Systems (NIPS), 2013, pp. 1–9.

[59] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 28, 2014.

[60] M. Nickel, L. Rosasco, T. Poggio, Holographic embeddings of knowledge graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30, 2016.

[61] A. Sheth, M. Gaur, U. Kursuncu, R. Wickramarachchi, Shades of knowledge-infused learning for enhancing deep learning, IEEE Internet Computing 23 (6) (2019) 54–63.

[62] I. B. Arpinar, U. Kursuncu, D. Achilov, Social media analytics to identify and counter islamist extremism: Systematic detection, evaluation, and challenging of extremist narratives online, in: 2016 International Conference on Collaboration Technologies and Systems (CTS), IEEE, 2016, pp. 611–612.