On Predicting Behavioral Deterioration in Online Discussion Forums

Jean Marie Tshimula,[1] Belkacem Chikhaoui,[1,2] Shengrui Wang[1]
[1]Prospectus Lab, Université de Sherbrooke, QC J1K 2R1, Canada
[2]LICEF Research Center, Université TÉLUQ, QC H2S 3L5, Canada
{kabj2801, shengrui.wang}@usherbrooke.ca, belkacem.chikhaoui@teluq.ca

***Abstract*—Early detection of behavioral deterioration can be of great importance in preventing individuals' misbehavior from escalating in severity. This paper addresses the problem of behavioral deterioration in the context of online discussion forums. We propose a novel method that builds behavioral sequences from temporal information to gain a better understanding of behaviors exhibited by forum members, and then explores n-gram features to predict behavioral deterioration from consecutive combinations of sequential patterns corresponding to misbehavior. We conduct extensive experiments using real-world datasets and demonstrate the ability of our method to predict behavioral deterioration with a high degree of accuracy, as evaluated by F-1 scores. Our quantitative analysis of the model's performance yields F-1 scores of over 0.7. Specifically, we find that the best-performing model is linear SVM, with an average F-1 score of 0.74. Some future research avenues are proposed.**

***Index Terms*—misbehavior, behavioral sequences, deterioration**

I. Introduction

The advent of online forums has revolutionized the speed of world connectivity, real-time information sharing, information discovery, real-time news, and instant communication, and creates new possibilities for investigating user behaviors through their digital footprints. Online forums aim to nurture social behavior, a sense of community and affinity relationships among individuals [40], [41]. Increasingly, however, they are having the opposite effect, due to a rising tide of deviations and deliberate provocations. While some people show common sense, tolerance and respect for the views of other forum members, others manifest intransigent attitudes and engage in misbehavior that harms the community and adversely affects the equanimity of forum members. The safety, usability, and reliability of online discussion forums may thus be compromised due to the prevalence of abuse and misbehavior expressed in various ways, such as videos, pictures, taunting emoticons and comments, to just name a few. In this paper, we limit our investigation to textual data and assemble different classes of temporal behavior displayed by individuals into more interpretable sequences.

Misbehavior may refer to disruptive acts characterized by covert or overt hostility and intentional aggression towards others [4], [15], [20], [21], [24], [30]. There is substantial evidence that the display of aggressive emotions is a valid predictor of risk factors for violence [1]. People who engage in misbehavior may severely transgress against social norms and social expectations for a particular environment, including full participation, right to safety and privacy, right to freedom of opinion and expression, decency, etc. Covert hostility can be expressed in one-to-one or one-to-many communication, whereas overt hostility can be voiced in online forums [24], [38]. It should be noted that misbehavior includes but is not limited to abusive and offensive language, threats, hate speech, cyberbullying, and race and gender discrimination [3], [6]. Waseem *et al.* (2017) [45] studied how these behaviors are related and proposed a typology that captures the similarities and differences among them. This provides ground truth for predicting future behavior with sufficient certainty.

Recent research has reported descriptive statistics on the number of victims of misbehavior. Kumar *et al.* (2017) [23] found that 40% of Internet users had experienced cyberbullying. Blumenfeld and Cooper (2000) [3] reported that 54% of LGBT youth had been cyberbullied. Li (2007) [27] found that nearly 54% of students were victims of traditional bullying and over a quarter of them had been cyberbullied. Additionally, their study found that roughly 60% of cybervictims were female and 39% were male. Waldman and Verga (2016) [43] put forward that 90% of terrorist activities on the Internet are conducted within online social networks. Some instances of misbehavior may initially have small statistical effects, but their persistent accumulation may subsequently have major and devastating consequences. Persistent misbehavior is a proven risk factor for a number of serious problems. For example, some victims of cyberbullying are more likely to self-harm, engage in suicidal behavior [22], and experience some unpleasant aftermaths, including psychological and anxiety disorders [7], [21], [28], [46]; others even commit suicide [18].

Evidence from the research discussed above shows a tremendous need for efficient approaches capable of preemptively detecting misbehavior as early as possible. In the absence of such approaches, misbehavior can escalate to violent behavior when the perpetrators constantly harm other forum members and do not get sanctioned for their misdeeds. Violent behavior may thus be considered as the endpoint on a continuum of behavioral deterioration [11]. Behavioral deterioration may occur suddenly or slowly, depending upon the pace at which perpetrators cause harm. Deterioration may be defined in many ways, and regardless of the definition, it is difficult to measure. More specifically, we define deterioration

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

as the accumulation of misbehavior.

The detection of misbehavior can be quite challenging and complex, for several practical reasons. Different people may have different ways of expressing the same misbehavior: for instance, masked pejorative terms, more subtle biases, coded messages and/or figures of speech (such as metaphor) may be used to misrepresent disparate impact [6], [37]. Recently, Mozafari *et al.* (2019) [31] introduced a BERT-based misbehavior classifier. This system suggests new fine-tuning strategies to investigate the effect of different layers of BERT and shows the ability to take contextual information into account, capture various ways in which misbehavior is expressed, and classify misbehavior classes more efficiently. In this paper, we resort to this model for building behavioral sequences from temporal behaviors exhibited by forum members in order to predict behavioral deterioration. To the best of our knowledge, our paper is the first to address the problem of behavioral deterioration in the context of online discussion forums.

Specifically, the key contributions of this paper can be summarized as follows:

- We first introduce a formal definition of the problem of behavioral deterioration.
- We then propose a method that constructs behavioral sequences from consecutive combinations of misbehavior classes and explores n-gram features to gain a better understanding of behavior exhibited by forum members and predict behavioral deterioration over time.
- We conduct extensive experiments using two publicly available datasets to validate the behavioral deterioration prediction. Our method is conceptually simple and highly interpretable.

The remainder of this paper is organized as follows. In Section II, we discuss some related work and the rationale for detecting signals relevant to deterioration. Section III describes the proposed method and the feature set extracted to train predictive models with alternative combinations of feature sets. We present experiments in Section IV. Section V is devoted to the discussion of our outcomes and the limitations of the study. Finally, we present our conclusions and propose future research directions in Section VI.

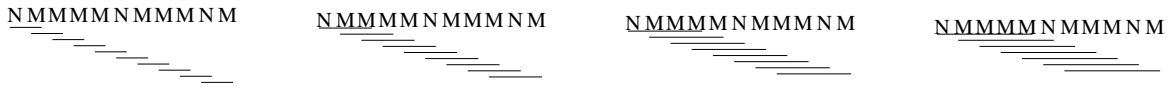
II. Related Work

Topic-based user behavior. Gong and Wang (2018) [14] introduced a holistic user behavior modeling approach to understand user intentions, relying on both sentiment and social network analysis to collect behavior patterns for each user. They developed a probabilistic generative model incorporating two learning tasks—opinionated content modeling and social network structure modeling—to recognize user preferences and their relatedness, respectively. In the first task, logistic regression is utilized to map sentiment polarity from textual content generated by a statistical language model based on a v-dimensional multinomial distribution over the vocabulary (v denotes the vocabulary size). In the second, a stochastic block model is employed to capture the relatedness among users. Wang *et al.* (2016) [44] explored the first task and proposed an unsupervised neural-network-based model to learn linguistic descriptors for the user's behavior over time. The method discovers linguistic dissimilarities that correlate with user activity levels and community clustering. While correlation does not imply causation, Aumayr and Hayes (2016) [2] sought to depict the correlation between clustered behaviors and three predefined topic properties (accessibility, sociability, and controversy). Their rationale was to present the effects that certain sorts of topics may have on user behavior, although the cluster categories were manually labeled to make the dendrogram more explicit. Furthermore, user behaviors were drawn from topics that they participated in rather than from opinions they expressed in the forum. We assume that this may result in failure to capture some signals that could be relevant to deterioration.

Hassan *et al.* (2010) [16] introduced a method for detecting the attitude of users towards others. Their approach involves training a supervised Markov model of the lexical item, part-of-speech tags, and dependency patterns to build a model capable of identifying sentences with and without attitude. On similar lines, Zhai *et al.* (2011) [47] proposed an unsupervised approach based on the evaluation of opinion sentences to remove those which contain emotional statements, personal attacks and opinions that do not express positive views about the discussion topics. Zhang *et al.* (2018) [48] detected early signs of conversational failures, such as harassment and personal attacks. More recently, Cliche (2017) [5] introduced a deep-learning-based classifier to tackle sentiment analysis issues. Their classifier leverages a large quantity of unlabeled information, using 100 million unlabeled tweets to pre-train word embeddings via distant supervision before applying convolutional neural networks and an attention-based biLSTM approach for classifying noisy positive and noisy negative tweets. We note some limitations of the aforementioned research, including the inability to verify whether individuals keep expressing opinions with or without attitude over time. In contrast to these studies, we examine temporal behaviors exhibited by forum members and assemble them in behavioral sequences to predict whether their behavior is affable or tends to deteriorate.

Zhao *et al.* (2015) [49] proposed a behavioral factorization (BF) method to model behaviors of each user based on topic interests derived from publishing signals such as posts, shares, likes, etc. BF learns a latent embedding model by factoring matrices split into behaviors (behavior-non-specific user-topic, single behavior-specific user-topic, and combined behavior-specific user-topic matrices) and then builds user topic profiles for various behavior types using the latent embedding space. The limitation of this work is that it draws solely on discussion topics addressed by forum members and does not regard different types of behavior they displayed in their posts.

Malicious and aggressive behavior. Cheng *et al.* (2015) [4] detected users engaged in antisocial behavior that negatively impinges on other users and causes harm to the community, and predicted whether some users would be banned from the



(a) Feature based on 2-grams (b) Feature based on 3-grams (c) Feature based on 4-grams (d) Feature-based on 5-grams

Fig. 1: Feature extraction process used to capture deterioration patterns within behavioral sequences (BS). Patterns are extracted based on all possible n-grams in BS from left-to-right. This allows us to better discern accumulations of behavior classes.

community based on their overall activities. Specifically, they compared the activities of users who have been banned in the past with those who have never been banned. To this end, the model deals with user posts, including data from features that allow users to upvote, downvote, report a post, etc. One limitation of this work is that the model relies more heavily on other features than on user posts to identify whether reported posts contain unpleasant statements. A post may be reported for the use of offensive language although the content of the post does not justify the accusations. The study does not address such cases. We believe that the model's failure to deal with post content is a shortcoming, as relying only on abuse-report-based features may be misleading to some extent.

Razavi *et al.* (2010) [35] reported work on multi-level classifiers enhanced by an Insulting or Abusive Language Dictionary (IALD) they developed to detect offensive language in text messages. Two rule-based auxiliary tools are proposed. One is the last level of the classifiers and the other is utilized for constructing patterns out of the IALD. Several solutions to the problems they address have been put forward in the literature, in particular for detecting cyberbullying, hate speech and offensive language in online communities [8], [17], [25], [29], [34], [36], [42], [45], [48]. In contrast to these studies, Mozafari *et al.* (2019) [31] proposed a BERT-based misbehavior classifier which outperforms several best-performing misbehavior classification techniques and understands and captures various ways in which misbehavior is expressed. We use this classifier [31] to construct behavioral sequences from temporal behaviors exhibited by individuals in order to predict behavioral deterioration.

III. Model

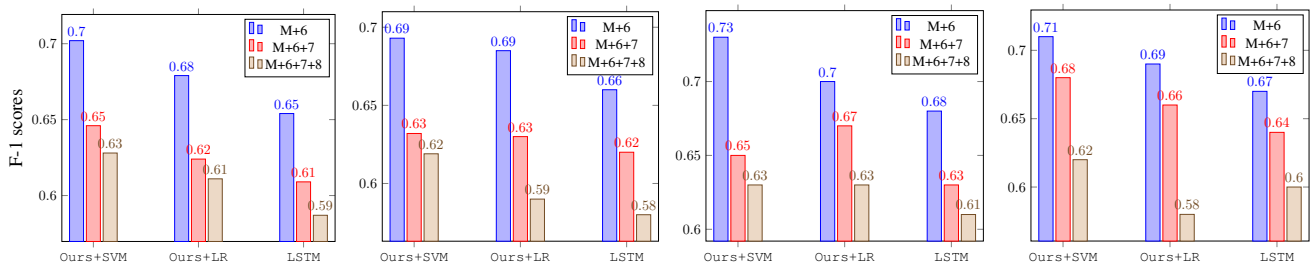
To illustrate our model, we utilize some simple notation. Let $S{=}\{s_1, s_2, \ldots, s_K\}$ denote a sequence of K sentences in a forum, $F{=}\{f_1, f_2, \ldots, f_H\}$ be a set of forum members, and $\alpha_S^{f_i}$ represent the set of sentences written by f_i in S, where $i{\in}\{1, \ldots, H\}$. A forum member participates in the discussion if there is an l such that $1{\leq}l{\leq}K$ and $s_l{\in}\,\alpha_S^{f_i}$. We assume that each such s_l is annotated beforehand in order to capture different types of behavior exhibited by f_i and facilitate behavior classification. Let $\beta_B^{f_i}{=}\{B_1, B_2, \ldots, B_T\}$ be the set of behavioral sequences exhibited by each forum member f_i, where $B_t{=}\{b_1^t, b_2^t, \ldots, b_m^t\}$ and $t{\in}\{1, \ldots, T\}$. Specifically, the sequence B_t represents the concatenation of all behavior label classes b_j^t exhibited by f_i in the period t, $\forall j{\in}\{1, \ldots, m\}$ and $b_j^t{\in}\{\texttt{N}, \texttt{M}\}$. The classes `N` and `M` designate normal behavior and misbehavior, respectively. It should be noted that the b_j^t are derived using a classifier.

To perform behavior classification, we use the BERT-based misbehavior classifier introduced in [31]. Fundamentally, BERT is a recent Transformer-based pre-trained contextualized embedding model extendable to a classification model with an additional output layer [9], [31]. It has yielded state-of-the-art results on numerous benchmarks, including text classification and language inference, without substantial task-specific modifications. The rationale behind the BERT-based misbehavior classifier [31] is that it exploits new fine-tuning strategies to capture different levels of syntactic and semantic information, and this enables it to consider tiny details in texts and to perceive different ways in which misbehavior is expressed. The contributions of this method are briefly discussed in [31].

Suppose that f_1 exhibits the behavior sequence `NMMMMNMMMNM` in the period t. The period is the interval of time elapsed between two different timestamps. We assume that deterioration cues can be observed from the accumulation of misbehavior classes.

To explore behavioral sequences, we design character n-gram features in order to capture signals that are potentially relevant to deterioration. The n-gram features with a pair of values (h_k, v_k) are extracted as input signals to be fed to a classifier. Specifically, h_k represents the n-gram feature k and v_k denotes the count of the feature within behavioral sequences. The n-grams can be generated by sliding a window of length n over the sequence B_t. Figure 1 illustrates how n-gram features can be extracted from behavioral sequences. For instance, the features extracted from the behavioral sequence above can be presented as follows: 2-grams {(`NM`, 3), (`MN`, 2), (`MM`, 5)}, 3-grams {(`MMM`, 3), (`NMM`, 2), (`MMN`, 2), (`MNM`, 2)}, 4-grams {(`MMMN`, 2), (`NMMM`, 2), (`MMNM`, 2), (`MMMM`, 1), (`MNMM`, 1)} and 5-grams {(`MMMNM`, 2), (`NMMMM`, 1), (`MMMMN`, 1), (`MMNMM`, 1), (`MNMMM`, 1), (`NMMMN`, 1)}.

To classify behavioral deterioration, we design four different features using n-grams of order 2, 3, 4 and 5, respectively (Figure 1). We use the constructed features to train linear support vector machines (SVM) and logistic regression (LR) classifiers. Basically, we label n-grams that support the accumulation of misbehavior classes as `Deterioration` and other n-grams as `Non-deterioration`. It should be noted that 4- and 5-grams which do not fully support the accumulation of misbehavior classes are treated differently. We consider them as full-fledged behavioral sequences and investigate the



(a) Deterioration (b) Non-deterioration (c) Deterioration (d) Non-deterioration

Fig. 2: Results of behavioral deterioration prediction by adding extra features to the main model. Specifically, M+6 means that we add 6-grams to the initially-built model, M+6+7 stands for 6- and 7-grams and M+6+7+8 means that we add 6-, 7- and 8-grams. We predict (a) and (b) on HatebaseTwitter and (c) and (d) on TRAC.

trend of their sub-2-grams by applying the same logic as in Figure 1(a). The choice of sub-2-grams is arbitrary. The principal reason for exploring sub-2-grams is to better track the momentum of the accumulation of different behavior classes and discover deterioration patterns. We label these 4- and 5-grams as Deterioration based on whether the majority of the sub-2-grams they contain support the accumulation of misbehavior classes. For instance, MNMM comprises {(MN, 1), (NM, 1), (MM, 1)}; NMMM, {(NM, 1), (MM, 2)}; MMMMN, {(MM, 3), (MN, 1)}; NMMMM, {(NM, 1), (MM, 3)}; and NMMMN comprises {(NM, 1), (MM, 2), (MN, 1)}. We, therefore, label them as follows: {NMMM, MMMMN, NMMMM} as Deterioration and {MNMM, NMMMN} as Non-deterioration.

IV. Experimental Setup

To empirically evaluate our method, we conducted experiments using two publicly available online discussion datasets: HatebaseTwitter [8] and TRAC [24].

Datasets. HatebaseTwitter is a collection of 24,802 tweets and contains three labels: hate, offensive, and neither. TRAC contains 15,869 Facebook comments labeled as overtly aggressive, covertly aggressive, and non-aggressive. To classify the class labels of experimental datasets, we applied the BERT-based misbehavior classifier [31]. This method outperforms [8] and [45] and yields accuracies of 96.2% and 94.8% on HatebaseTwitter and TRAC, respectively (versus 90% for [8] on HatebaseTwitter, and 80% and 89% for [45] on TRAC and HatebaseTwitter). We, therefore, took the predicted classes produced by [31] to design behavioral sequences on a weekly basis: i.e., each sequence represents behaviors exhibited by an online forum member in the course of the week. The choice of the period over which to form the behavioral sequence is arbitrary and depends on how one wants to learn the deterioration distribution. To better explore sequence variation and follow deterioration cues, we chose to simplify the sequence by converting all misbehavior-related classes into "M" and the normal behavior class into "N". The major reason for using binary classes is to explore the behavioral sequences with a small number of object types in order to examine them thoroughly. We, therefore, utilized the designed features and the two classifiers for experimental settings, as mentioned in Section III.

TABLE I: Results of behavioral deterioration prediction. Bold font indicates the best results for each class label.

	Class	HatebaseTwitter	TRAC
Ours+SVM	Deterioration	**0.722**	**0.785**
	Non-deterioration	0.718	0.749
Ours+LR	Deterioration	0.72	0.761
	Non-deterioration	**0.719**	**0.758**
LSTM	Deterioration	0.719	0.737
	Non-deterioration	0.716	0.733

Model evaluation. To evaluate the performance of our model, we used 10-fold cross-validation to split our training and testing sets. We computed F-1 scores to measure the accuracy of our classifiers and quantitatively compared them with the baseline. We used long short-term memory (LSTM) [19] as the baseline since it deals very well with long sequences and captures long-term dependencies. Note that we did not find an existing approach for detecting behavioral deterioration in the context of online forums.

V. Results and Discussion

We show that quantifiable signals relevant to accumulations of misbehavior classes can be used for behavioral deterioration prediction. Table I presents the performance results of our method and the baseline. We observe that the F-1 scores for our classifiers and LSTM are significantly higher and show the ability to predict behavioral deterioration, with F-1 scores of over 0.7 for both classes. All classifiers showed significantly better results for the class Deterioration. Note that our method achieved higher F-1 scores on both datasets. The results of LSTM on HatebaseTwitter are not far behind, while on TRAC the differences widen by a considerable margin for both classes, especially evident in the values 0.048 and 0.024 for the class Deterioration with Ours+SVM and Ours+LR, respectively. It should be noted that Ours+SVM was the best-performing classifier, yielding an average F-1 score of 0.74, and Ours+SVM and Ours+LR achieve ap-

proximately the same results on HatebaseTwitter. Additionally, we note that Ours+LR performs in more balanced ways and remark that the differences between its predicted class labels are smaller than those yielded by Ours+SVM: ($0.001 < 0.004$) on HatebaseTwitter and ($0.003 < 0.036$) on TRAC.

To validate the performance of our models in generating deterioration estimates adequate to provide a good solution for particular individuals, some ground truth information is required. However, ground truth information to capture the prediction accuracy of behavioral deterioration is scarce and constitutes a challenging problem that has not been addressed in the context of online forums. It should be noted that the lack of ground truth information does not affect the generalizability of the findings and model performance, since the results stem directly from observed accumulations of behaviors exhibited by individuals in the discussion forum; and this makes intuitive sense. The number of features, as well as the number of elements in each n-gram, may be arbitrary and depend heavily on the average length of behavioral sequences. To explore the effect of the number of features on model performance, we extend the initially-built model by including in it some supplementary features to examine deterioration patterns within behavioral sequences. It should be recalled that the average length of the set of behavioral sequences that we constructed above is 9. Consequently, we add to the initially-built model a feature extracted on 6-grams (M+6); two features extracted on 6- and 7-grams, respectively (M+6+7); and three features extracted on 6-, 7- and 8-grams, respectively (M+6+7+8). We treated differently 6-, 7- and 8-grams which do not fully support the accumulation of misbehavior classes by applying the same logic as for 4- and 5-grams, as described in §III.

Figure 2 presents the results of behavioral deterioration prediction with additional features. We report that the average performances yielded by our models exceed 0.6; that is, (0.691, 0.635, 0.619) for Deterioration and (0.689, 0.631, 0.605) for Non-Deterioration with M+6, M+6+7, and M+6+7+8, respectively, on HatebaseTwitter; and (0.715, 0.66, 0.63) for Deterioration and (0.7, 0.67, 0.605) for Non-Deterioration with M+6, M+6+7, and M+6+7+8, respectively, on TRAC. Our models achieved better results than LSTM on the two experimental datasets. We observe that the model performance decreases when the number of features increases. To examine deterioration patterns more closely, we suggest constructing a model based on z-1 features if the average length of overall behavioral sequences corresponds to z ($z > 2$). Following this logic, the model is supposed to utilize the feature sets varying from 2-grams to (z-1)-grams. We assume that this renders it possible to extract longer accumulations of behavior classes to investigate deterioration patterns on various facets. Beyond monitoring accumulations of behavior classes to extract feature sets, we face challenges in defining threshold values (or early warning scores) to determine whether a set of behavioral sequences for individuals tends toward deterioration or not. Such scores could allow the establishment of different degrees of deterioration in order to facilitate more effective monitoring of the trajectory of behavioral deterioration. With thresholds fixed, we can identify deterioration at a sufficiently early stage to prevent significant further deterioration [12] and examine an individual's mental state and personality traits [13], [33].

Our results provide strong evidence that we can predict behavioral deterioration with an accuracy exceeding 0.6 (Table I and Figure 2), a resolution that is likely fine-grained enough for various experimental datasets. Significant signals relevant to deterioration remain to be uncovered and understood within behavioral sequences, including *(i)* examining correlations between language use of individuals for which behavior sequences comprise accumulations of behavior classes that indicate signals relevant to deterioration; *(ii)* analyzing personality traits to understand whether deterioration occurs under the effects of the topics addressed in the discussion forum, mental health conditions or some other factors and *(iii)* understanding the impact of some personal concerns (such as work, money, religion, death, etc.) on behavioral deterioration; i.e. constructing a holistic model to explain the deterioration in conjunction with several factors [10], [26]. Developing these algorithms and evaluating them is a promising direction for future research.

VI. Conclusion

We present a method that constructs behavioral sequences from forum members' temporal activities and behaviors, to predict behavioral deterioration. We explore deterioration patterns from consecutive combinations of behavior classes corresponding to misbehavior, utilizing two publicly available datasets. We achieve F-1 scores as high as 0.7 with the initially-built model and 0.6 when alternative features are added to the initially-built model. Our method provides a straightforward way to obtain signals relevant to deterioration without involving other contributing factors, such as an individual's mental state, personality traits, and affinity relationships [41]. Some of these opportunities are discussed in Section V; i.e., fixing deterioration threshold and building a holistic model for determining the magnitude of deterioration.

This problem leaves room for future research. In the future, we aim to add multimodal analysis and investigate behavioral sequences without converting misbehavior-related classes into a single class category. Furthermore, we would like to work on measuring the distance and similarity between multiple behavioral sequences,[1] predicting affinity relationships between individuals who exhibit deteriorating behaviors, identifying among these individuals those who seem to foment misbehavior within the online discussion forums, and assessing the likelihood that their affinity may evolve and the risks they may represent.

References

[1] C.A. Anderson, and B.J. Bushman, "Human aggression," Annual Review of Psychology, 53:27–51, 2002.

[1]To this end, we plan to address the behavioral sequences as biological sequences in order to apply sequence alignment-based algorithms such as Needleman–Wunsch [32], Smith–Waterman [39], etc.

[2] E. Aumayr, and C. Hayes, "On the correlation between topic and user behaviour in online communities," In Proceedings of the Tenth ICWSM, pp. 531–534, 2016.

[3] W.J. Blumenfeld, and R.M. Cooper, "Lgbt and allied youth responses to cyberbullying: Policy implications," International Journal of Critical Pedagogy, 3(1):114–133, 2000.

[4] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Antisocial behavior in online discussion communities," In Proceedings of the Ninth ICWSM, pp. 61–70, 2015.

[5] M. Cliche, "BBtwtr at SemEval-2017 Task 4: Twitter sentiment analysis with CNNs and LSTMs," In Proceedings of the 11th International Workshop on Semantic Evaluations, pp. 573–580, 2017.

[6] C.H. Coleman, "The disparate impact argument reconsidered: Making room for justice in the assisted suicide debate," The Journal of Law, Medicine and Ethics, 30(1):17–23,2002

[7] H. Cowie, "Cyberbullying and its impact on young people's emotional health and well-being," The Psychiatrist, 37(5):167–170, 2013.

[8] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," In Proceedings of the Eleventh ICWSM, pp. 512–515, 2017.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," In Proceedings of NAACL, pp. 4171–4186, 2019.

[10] L.H. Dewa, E. Cecil, L. Eastwood, A. Darzi, and P. Aylin, "Indicators of deterioration in young adults with serious mental illness: a systematic review protocol," Systematic Reviews, 7:123, 2018.

[11] F. Fluttert, B. Van Meijel, M. Van Leeuwen, S. Bjørkly, H. Nijman, and M. Grypdonck, "The development of the forensic early warning signs of aggression inventory: preliminary findings: Toward a better management of inpatient aggression," Arch Psychiatr Nurs., 25(2):129–137, 2011.

[12] C. Gaskin, and G. Dagley, "Recognising signs of deterioration in a person's mental state," Sydney: Australian Commission on Safety and Quality in Health Care, 2018.

[13] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," In Proceedings of 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, pp. 149–156, 2011.

[14] L. Gong, and H. Wang, "When sentiment analysis meets social network: A holistic user behavior modeling in opinionated data," In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, pp. 1455–1464, 2018.

[15] M.A. Hamilton, "Verbal aggression: Understanding the psychological antecedents and social consequences," Journal of Language and Social Psychology, 31(1):5–12, 2011.

[16] A. Hassan, V. Qazvinian, and D. Radev, "What's with the attitude? Identifying sentences with attitude in online discussions," In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1245–1255, 2010.

[17] C.V. Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, G.D. Pauw, W. Daelemans, and V. Hoste, "Automatic detection of cyberbullying in social media text," PLoS ONE, 13(10), 2018.

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

[19] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," Neural Computation, 9(8):1735–1780, 1997.

[20] I. Hsieh, and Y.Y. Chen, "Determinants of aggressive behavior: Interactive effects of emotional regulation and inhibitory control," PLoS One, 12(4):e0175651, 2017.

[21] L.R. Huesmann, and L.D. Taylor, "The role of media violence in violent behavior," Annual Review of Public Health, 27:393–415, 2006.

[22] A. John, A.C. Glendenning, A. Marchant, P. Montgomery, A. Stewart, S. Wood, K. Lloyd, and K. Hawton, "Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review," Journal of Medical Internet Research, 20(4):e129, 2018.

[23] S. Kumar, J. Cheng, and J. Leskovec, "Antisocial behavior on the web: Characterization and detection," In Proceedings of the 26th International Conference on World Wide Web, pp. 947–950, 2017.

[24] R. Kumar, A.K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pp. 1–11, 2018.

[25] I. Kwok, and Y. Wang, "Locate the hate: detecting tweets against blacks," In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, pp. 1621–1622, 2013.

[26] J.E. Lamond, R.D. Joseph, and D.G. Proverbs, "An exploration of factors affecting the long term psychological impact and deterioration of mental health in flooded households," Environmental Research, 140:325–334, 2015.

[27] Q. Li, "New bottle but old wine: A research of cyberbullying in schools," Computers in Human Behavior, 23(4):1777–1791, 2007.

[28] J. Lindert, "Cyber-bullying and it its impact on mental health," European Journal of Public Health, 27(3), 2017.

[29] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, "Hate speech detection: Challenges and solutions," PLoS ONE, 14(8):e0221152, 2019

[30] M. Mengu, and S. Mengu, "Violence and social media," Athens Journal of Mass Media and Communications, 1:211–228, 2015.

[31] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," Computational Intelligence, 881:928–940, 2019.

[32] S.B. Needleman, and C.D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," Journal of Molecular Biology, 48(3):443–53, 1970.

[33] B. Plank, and D. Hovy, "Personality traits on twitter—or—how to get 1,500 personality tests in a week," In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 92–98, 2015.

[34] R.I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, and S. Mishra, "Scalable and timely detection of cyberbullying in online social networks," In Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 1738–1747, 2018.

[35] A.H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, "Offensive language detection using multi-level classification," In Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence, pp. 16–27, 2010.

[36] S. Salawu, Y. He, and J. Lumsden, "Approaches to automated detection of cyberbullying: A survey," IEEE Transactions on Affective Computing, pp. 1–20, 2017.

[37] M. Selmi, "Was the disparate impact theory a mistake?," UCLA Law Review, 53(3):701–782, 2006.

[38] M. Seufert, T. Hoßfeld, A. Schwind, V. Burger, and P. Tran-Gia, "Group-based communication in whatsapp," In Proceedings of 2016 IFIP Networking Conference and Workshops, pp. 536-541, 2016.

[39] T.F. Smith, and M.S. Waterman, "Identification of common molecular subsequences," Journal of Molecular Biology, 147(1):195–197, 1981.

[40] J.M. Tshimula, B. Chikhaoui, and S. Wang, "Har-search: A method to discover hidden affinity relationships in online communities," In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 176–183, 2019.

[41] J.M. Tshimula, B. Chikhaoui, and S. Wang, "A New Approach for Affinity Relationship Discovery in Online Forums," Social Network Analysis and Mining, 10, 40, 2020.

[42] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," In Proceedings of the 21st ACM international conference on Information and knowledge, pp. 1980–1984, 2012.

[43] S. Waldman, and S. Verga, "Countering violent extremism on social media," In Defence Research and Development Canada, 2016.

[44] A. Wang, W.L. Hamilton, and J. Leskovec, "Learning linguistic descriptors of user roles in online communities," In Proceedings of the First Workshop on NLP and Computational Social Science, pp. 76–85, 2016.

[45] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understand abuse: A typology of abusive language detection subtasks," In Proceedings of the First Workshop on Abusive Language Online, pp. 78–84, 2017.

[46] H. Yenala, A. Jhanwar, M.K. Chinnakotla, and J. Goyal, "Deep learning for detecting inappropriate content in text," International Journal of Data Science and Analytics, 6(4):273–286, 2018.

[47] Z. Zhai, B. Liu, L. Zhang, H. Xu, and P. Jia, "Identifying evaluative sentences in online discussions," In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, pp. 933–938, 2011.

[48] J. Zhang, J. Chang, C. Danescu-Niculescu-Mizil, L. Dixon, Y. Hua, D. Taraborelli, and N. Thain, "Conversations gone awry: Detecting early signs of conversational failure," In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 1350–1361, 2018.

[49] Z. Zhao, Z. Cheng, L. Hong, and E.D. Chi, "Improving user topic interest profiles by behavior factorization," In Proceedings of the 24th International Conference on World Wide Web, pp. 1406–1416, 2015.