Identifying Latent Toxic Features on YouTube Using Non-negative Matrix Factorization

Adewale Obadimu
Department of Computer Science
University of Arkansas at Little Rock
Little Rock, USA
Email: amobadimu@ualr.edu

Esther Mead
Department of Information Science
University of Arkansas at Little Rock
Little Rock, USA
Email: elmead@ualr.edu

Nitin Agarwal
Department of Information Science
University of Arkansas at Little Rock
Little Rock, USA
Email: nxagarwal@ualr.edu

***Abstract*—Toxic behavior, in its various forms, often disrupts constructive discussions in online communities. The proliferation of smart devices and mobile applications has further exacerbated these nefarious acts on various social media platforms. Largely, toxic behavior is regulated by human moderators employed by the platform operators. However, given the volume and speed of content posted on online platforms, identifying and deterring these behaviors remains challenging. In this study, we propose a Non-negative Matrix Factorization (NMF) technique for predicting commenter toxicity on YouTube. We utilized the YouTube Data API to collect data from the Cable News Network (CNN) channel on YouTube. Our final dataset consists of 144 videos, 243,344 commenters, and 421,924 comments. We then utilized Google's Perspective API to assign a toxicity score to each comment. We used the resultant dataset to create a commenter toxicity score prediction model. We tested our proposed NMF model against other popular prediction methods, comparing speed of model execution and the common Root-Mean-Square-Error (RMSE) accuracy metric. This work sets the stage for a richer, more detailed analysis of toxicity on various online social media networks.**

Keywords—Toxicity; Tonality Analysis; YouTube; Social Media; Language Model.

I. INTRODUCTION

The advent of computer-mediated forms of interactions has posed several conceptual and practical challenges [1]. With emergent norms and conventions, the Web, more than any other medium, has offered lowered communication thresholds and a broadened geographical scope of human interactions [2]. However, despite the myriad advantages of utilizing this medium to connect to like-minded individuals, a consensus is emerging suggesting the presence of malicious actors, otherwise known as trolls [3]. These actors (hereafter referred to as *toxic users*) thrive on disrupting the norms of a given platform and causing emotional trauma to other users [4]. In this study, similar to extant literature, we give an operational definition of toxicity as "the usage of rude, disrespectful, or unreasonable language that will likely provoke or make another user leave a discussion" [3][5][6][7]. Therefore, in this regard, toxicity analysis is different from sentiment analysis, which is the attempt to assign sentiment scores of positive, neutral, and negative to text data.

Social media was once perceived as a liberating platform but is now riddled with various forms of toxicity [5]. A report by the Pew Research Center indicated that 73% of adult Internet users have seen someone harassed online, and 40% have experienced it personally [3][8]. Another survey by Duggan [8] highlighted that 19% of teens reported that someone has written or posted malicious or embarrassing things about them on social networking sites. Due to the growing concerns about the impact of online harassment, many platforms are taking several steps to curb this phenomenon [5][6][9][10]. For instance, on YouTube, a user can simply activate the safety mode to filter out offensive language [8]. Wikipedia has a policy of "Do not make personal attacks anywhere in Wikipedia" [5]. Likewise, platforms like CNN.com have moderators that reportedly remove over one in five comments that violate community guidelines on any given day [3]. The aforementioned are a few examples that highlight the negative impacts of toxic behavior on the community. Toxic behavior, if not curbed at the initial stage, can have a ripple effect. It can dissuade other people from joining a community by perceiving the community as a hostile environment [5].

According to Alexa, the web traffic monitoring service owned by Amazon, YouTube is the second most popular website globally with over 300 hours of videos uploaded every minute and 5 billion videos watched every single day [10]. While several studies have attested to the widespread manifestation of toxicity within comments on YouTube [5][11][12], the burden of access to such a large dataset has made a permanent extraction of toxic users challenging. Due to the extensive exploitation of these platforms by toxic users, automatic detection and extraction of toxicity has become a pressing need [9].

Despite the rich vein of academic research on identifying various forms of online toxicity, tackling these behaviors at scale remains surprisingly challenging [3]. This, along with the immensity of the amount of data and the speed with which the data is generated and shared, motivated us to propose a Non-negative Matrix Factorization (NMF) technique for unraveling latent toxic commenter features on YouTube. The intuition behind

 ISBN: 978-1-61208-757-3

using matrix factorization to solve this problem is that there should be some latent features that determine the toxicity of a commenter on a given video. For instance, two commenters may have higher toxicity on a video if they both dislike who or what the video is talking about, or perhaps, if they both mutually dislike the genre of the video. Hence, if we can automatically discover these latent features, we should be able to predict the toxicity of a certain commenter on a certain video. This work is novel in that the NMF approach has not previously been applied to this type of problem. The contribution of this work is to help understand the relationship between the impact of toxicity of a video on the comments, to enable us to predict the likely toxicity of a commenter based on their past history, and to allow us to determine what kind of comments a video will generate based on prior toxicity matrix. We selected the CNN news channel's "*must-see moments*" videos section as an experimental dataset because it is rich in various forms of behaviors.

The remainder of this paper is set out as follows. In Section 2, we present a theoretical formulation of the problem. Then, we give a brief review of extant literature that are most germane to our discussion in Section 3. Next, our methodology is described in Section 4, and the findings are discussed. Section 5 provides conclusions including the limitations of our work, and ideas for future work.

II. PROBLEM FORMULATION

We pose NMF as an optimization problem where, in addition to minimizing the reconstruction error of the commenter-video toxicity matrix, we also require that the factors capture prior knowledge as much as possible. The intuition behind using this approach is that there should be some latent features (characteristics that are not directly observed [30]) that determine the toxicity of a given commenter on a specific video. In trying to discover these latent features, we assume that the number of features would be smaller than the number of commenters and the number of videos. To validate the efficacy of this approach, we applied this method to a real-world YouTube dataset, making our work the first to conduct toxicity analyses using NMF on YouTube.

Given a set of commenters $C = \{c_1, \dots, c_N\}$ and a set of videos $V = \{v_1, \dots, v_M\}$, the toxicity expressed by these commenters on all the videos can be expressed in a toxicity matrix $T = [T_{c,v}]_{N \times M}$. In this matrix, $T_{c,v}$ represents the average toxicity of a commenter c on a video v and it is bounded in the range of $[0,1]$. Our objective in this study is as follows: Given a commenter $c \in C$ and a video $v \in V$ for which $T_{c,v}$ is unknown, predict the toxicity for c on video v using T. T is asymmetric and usually very sparse.

Let $P \in R^{K \times N}$ and $Q \in R^{K \times M}$ be latent commenter and video feature matrices, with column vectors P_c and Q_v representing K-dimensional commenter-specific and video-specific latent feature vectors of commenter c and video v, respectively. The resulting dot product ${P_c}^T Q_v$ captures the interaction between commenter c and video v. This product approximates commenter c's toxicity on video v, and it is denoted by $\hat{T}_{c,v}$ as shown in (1).

$$\hat{T}_{c,v} = {P_c}^T Q_v \quad (1)$$

To learn the latent feature vectors (P_c and Q_v), we minimize the regularized squared error (2) on the set of known toxicity using stochastic gradient descent.

$$min \sum_{(c,v)\in\Im} (T_{c,v} - \hat{T}_{c,v})^2 + \lambda(||(P_c|| + ||Q_v||) \quad (2)$$

Here, $\Im$ is the set of the (c, v) pairs for which $T_{c,v}$ is known (the training set). The conditional probability of the observed toxicity (3) is defined as:

$$p(T|P, Q, \sigma_T^2) = \prod_{c=1}^{N} \prod_{v=1}^{M} \left[N(T_{c,v}|({P_c}^T Q_v), \sigma_T^2\right] \quad (3)$$

where $N(x|\mu, \sigma_T^2)$ is a normal distribution with mean μ and variance σ^2.

III. RELATED WORK

This section discusses the two categories of related work that correspond to our research. The first category includes works that have attempted to identify toxicity on social media. The second category includes the works that have attempted to utilize matrix factorization techniques.

A. Identifying toxicity on Social Media

Threads of extant literature on antisocial behavior suggest that toxicity, in its various forms, oftentimes disrupts constructive discussions in an online community [3][5][12][13]. Several researchers have tried to identify and suggest ways to mitigate toxicity in a community [5][13]. Using data collected via crowdsourcing, Wulczyn et al. [5] employed machine learning techniques, such as linear regression and multilayer perceptron to analyze personal attacks on social media. A study by Martens et al. [6] utilized Natural Language Processing techniques to detect the emergence of undesired and unintended behavior in online multiplayer games. Research by Chen et al. [14] and Yin et al. [15] used a set of regular expressions, n-grams, and supervised learning techniques to detect abusive language. Sood et al. [16] combined lexical and parser features to identify offensive language

 ISBN: 978-1-61208-757-3

in comments extracted from a social news site. Davidson et al. [17] presented a dataset with three kinds of comments: hate speech, offensive but non-hateful speech, and neither. Hosseini et al. [18] demonstrated the vulnerability of most state-of-the-art toxicity detection to adversarial inputs. Despite the rich vein of academic work on toxicity detection, there is a need for systematic research that focuses on detecting, identifying, and categorizing toxicity at scale.

B. Matrix Factorization Techniques

Our work using NMF was heavily inspired by the need for an approach that scales to the huge amount of streaming data on YouTube. The idea behind matrix factorization is to decompose an interaction matrix into a product of two lower dimensionality rectangular matrices while minimizing the error associated with the decomposition [19]. This technique has been used extensively in recommendation systems to discover latent features underlying the interactions between users and items ratings [20][21]. Yang et al. [22] applied a matrix factorization technique for developing a model-based community detection algorithm that detects densely overlapping communities in a network. Ma et al. [20] utilized a probabilistic matrix factorization technique to solve the data sparsity and poor prediction accuracy problems by employing both users' social network information and rating records to perform recommendation. Zhao et al. [23] employed a matrix factorization method on each of the communities in a unidirectional social network. By advancing previous work, Jamali et al. [24] proposed a matrix factorization technique with trust propagation for recommendation systems. Chen et al. [25] proposed a novel social recommendation method that fuses user's social status with homophily using a matrix factorization technique. Peng et al. [21] proposed a social trust and segmentation-based matrix factorization recommendation algorithm. Ozer et al. [26] leveraged matrix factorization techniques to uncover political networks on Twitter. However, to the best of our knowledge, our work is the first to apply matrix factorization to unravel toxic features on YouTube.

IV. METHODOLOGY

Our methodology (Fig. 1) consists of three phases: 1) data collection and data processing; 2) data preparation and toxicity assignment; and, 3) matrix factorization of commenter-video toxicity matrix.

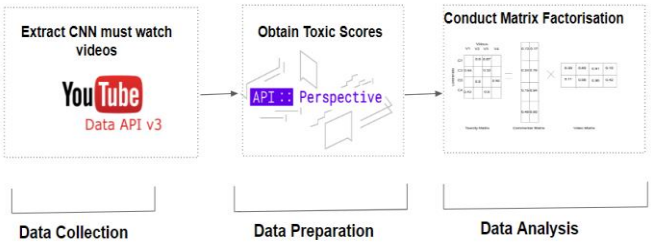

Figure 1. Research methodology.

A. Data Collection

To build our dataset for illustrating our proposed method, we first utilized Google's YouTube API to extract videos' and commenters' data from the "*must see moments"* video playlist from the CNN channel. To reduce "noise" in extracted data, several data processing steps were subsequently performed including data formatting, data standardization and data normalization using the Python programming language. Our final dataset consists of 144 videos, 243,344 commenters, and 421,924 comments.

B. Data Preparation

The next step in our methodology was to assign toxicity scores to each comment in the dataset. To accomplish this, we leveraged a classification tool called Perspective API [27], which was developed by Google's Project Jigsaw' and 'Counter Abuse Technology' teams (Table I). This model uses a Convolutional Neural Network (CNN) trained with word-vector inputs to determine whether a comment could be perceived as "toxic" to a discussion [27][28]. The API returns a probability score between 0 and 1, with higher values indicating a greater likelihood of the toxicity label. Table I is an excerpt of the resultant toxicity dataset where toxicity scores have been assigned to each comment.

TABLE I. CONVENIENCE SAMPLING OF FIVE (5) TOXIC COMMENTS IN OUR DATASET.

S/N	*Comment*	*Overall Toxicity*
1	SHUT UP YOU OLD FOOL !	0.97
2	dumba** liberal kids..#cnnsucks	0.74
3	JIM your a special kind of STUPID..	0.97
4	Brian, you are so dumb.	0.96
5	Stupid dumb gun lover.	0.78

C. Data Analysis

Before applying algorithms for toxicity prediction, we conducted some preliminary data analysis on our YouTube CNN video comments dataset. Fig. 2 shows the distribution of overall toxicity in our dataset. Although most of the comments (80%) were assigned a toxicity score of less than 0.5, a significant portion (20%) of the

 ISBN: 978-1-61208-757-3

comments were assigned a toxicity score of greater than 0.5 (over 84,000 comments). The mean toxicity score was about 0.27.

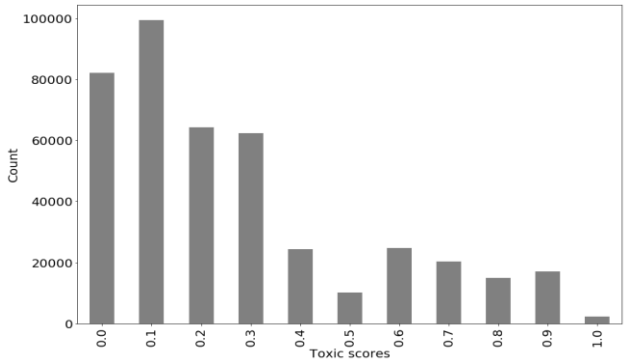

Figure 2. Distribution of overall toxicity.

Most of the toxic comments within the CNN dataset were categorized as being some type of “INSULT”—a negative comment towards a person or a group of people, followed closely by “THREAT”—an intention to inflict pain, injury, or violence against an individual or group (Fig. 3).

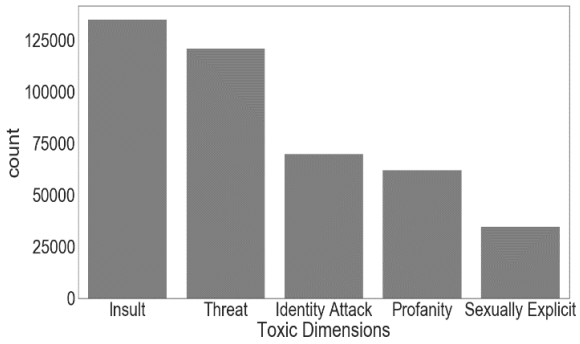

Figure 3. Distribution of toxicity by toxicity dimension.

Fig. 4 shows the box plots for each of the toxicity dimensions and reveals that the “INSULT” and “PROFANITY”—the usage of swear/curse words, or other obscene/profane language—had the widest spread with regard to the distribution of toxicity scores. However, compared to other toxicity dimensions, sexually explicit comments were the least severe.

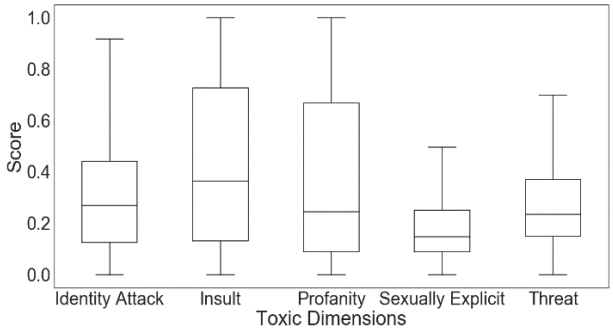

Figure 4. Box plot for each toxicity dimension.

Word clouds generated from the dataset reveal some of the content that relates to these five toxicity dimensions (Fig. 5-7). As can be seen in the content of each word cloud, some of the categories are more obvious than others; for example, the prevalence of terms such as, “kill” and “shooting” within the “THREAT” toxicity dimension (Fig. 6). Similarly, the prevalence of terms such as, "black", "white", "fake", "Democrat", and "Republican" within the "IDENTITY ATTACK" toxicity dimension (Fig. 7) give indications of why these comments were categorized as belonging to this specific toxicity dimension. Some of the word clouds that we generated could not be included in this work because of the obscenity, profanity, and sexually explicit terms that they revealed. These word clouds corresponded to the toxicity dimensions of "PROFANITY" and "SEXUALY EXPLICIT".

Figure 5. “INSULT” word cloud.

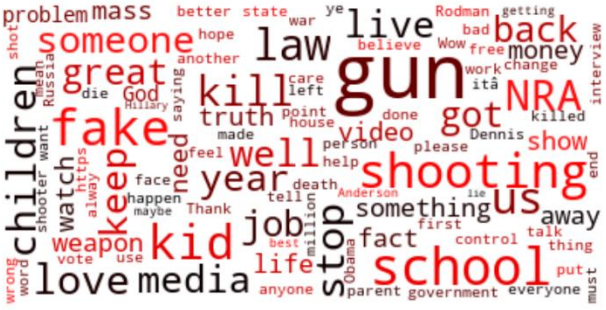

Figure 6. “THREAT” word cloud

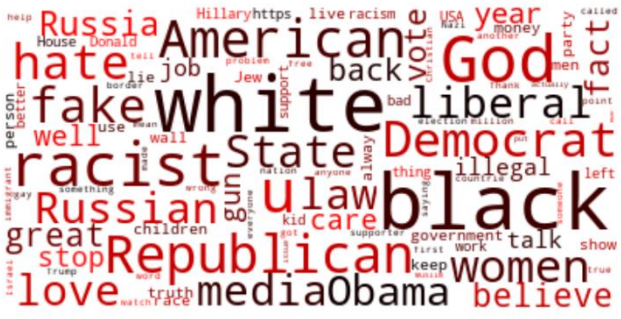

Figure 7. “IDENTITY ATTACK” word cloud

D. Evaluation of the approach

The next step in our methodology was to apply the algorithms for toxicity prediction and compare the results. The most commonly used metrics for assessing the effectiveness of predictive methods, such as the proposed NMF algorithm are the mean squared error and Root-Mean-Square-Error (RMSE), the latter having been used in the Netflix Prize [19]. RMSE is a measure that can be used to compare predictions against real data. The smaller the RMSE value, the better the model. The error was computed, and gradient descent was performed to minimize the error. We used a random 70/30 split of our dataset to create training and test sets and applied 5-fold cross-validation. Using *Surprise Python Scikit Package* [29], we compared the result of the NMF approach with other techniques: CoClustering (Collaborative filtering algorithm) and NormalPredictor (Algorithm based on the normal distribution of the training set) (Table II).

TABLE II. EVALUATING RMSE ON 5 SPLIT(S). SMALLER THE VALUE THE BETTER THE MODEL.

Algorithm	*Mean RSME*
NMF **Improved %**	**0.28** **17.85%**
CoClustering	0.33
NormalPredictor	0.34

Compared to CoClustering and NormalPredictor, the NMF approach performs better in terms of accuracy (having the lowest RMSE). Table III shows that the NMF approach outperformed NormalPredictor based on computation time, but not CoClustering. The mean computation time for NMF was 0.54 seconds, while NormalPredictor took 0.56 seconds and CoClustering took 0.45 seconds. This experiment was conducted on a machine with Intel(R) Xeon(R) CPU E7-8893 v4 @ 3.20GHz 3.19 GHz (4 processors) and 3.25 TB RAM. The complexity analysis indicates that our approach can be applied to very large datasets since it scales linearly with the number of observations.

V. CONCLUSION

In this study, we addressed the problem of identifying and predicting toxicity on online social media networks. We chose to focus on user comments posted on a sample of CNN videos posted on YouTube as a case study for illustrating our methodology. The challenges in addressing this problem include an ongoing issue of balancing freedom of expression with curtailing harmful content. The contribution of this paper is that it outlines a scalable methodology for first identifying toxicity within commenter text data posted on an online social media network and then predicting the toxicity levels of each commenter. We found that the proposed NMF-based approach performed better than some other techniques for predicting toxicity scores in terms of and accuracy and has the potential to perform better in terms of computation time. These findings demonstrate how the presence of toxicity in a set of text corpora can be identified, categorized, and measured, and how those metrics can be used for prediction. Our findings advance the research in this topic in that this analysis serves as a steppingstone in a long line of future work that can be done to better understand the origin, propagation and impact of toxicity on online social media networks. The general implications of this work are that by systematically assessing toxicity, this work sets the stage for developing a richer, more robust model for understanding the flow of toxicity on various online social media networks and for developing tools for toxicity control and prevention. There are a few limitations to this work, however. One is that it is challenging to model a cold-start commenter—a commenter that has never posted a comment—as new users will not have an initial toxicity score. Additionally, the commenter-video toxicity matrix can become very sparse with increasing volume of data thereby increasing the reconstruction error. Our immediate next steps will investigate possible solutions to address the limitations. Long term future work includes running experiments on multiple datasets and considering other variables to determine whether this can improve prediction accuracy. We anticipate applying the NMF method to data from other online social media networks. For instance, a sophisticated version of the NMF approach can be used to capture the bidirectional latent relations between user's toxicity preferences across domain through transfer learning. Future work should also attempt to model the spread of toxicity on online social media networks and determine whether these metrics and predictions can be used for developing preventive measures.

ACKNOWLEDGMENT

This research is funded in part by the U.S. Office of Naval Research (N00014 - 10 - 1 - 0091, N00014 - 14 - 1 - 0489, N00014-15-P-1187, N00014-16-1-2016, N00014 - 16 - 1- 2412, N00014-17-1-2605, N00014-17-1-2675, N00014-19-1-2336), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, and the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily

 ISBN: 978-1-61208-757-3

reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

REFERENCES

[1] M. Warschauer, "Computer-mediated collaborative learning: Theory and practice." The Modern Language Journal, vol. 81, no. 4, pp. 470-481, 1997.

[2] E. Eleanor, "Hyperbole over cyberspace: Self-presentation and social boundaries in Internet home pages and discourse," The Information Society, vol. 13, no. 4, pp. 297-327, 1997.

[3] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions," Proc. ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW17), ACM Press, Feb. 2017, pp. 1217–1230, doi: 10.1145/2998181.2998213.

[4] S. Lee and H. Kim, "Why people post benevolent and malicious comments online," Communications of the ACM, vol. 58, no. 11, pp. 74–79, 2015.

[5] E. Wulczyn, T. Nithum, and L. Dixon, "Ex machina: Personal attacks seen at scale," Proc. 26th International Conference on World Wide Web (WWW 2017) ACM, [month] 2017, pp. 1391-1399, ISBN: 978-1-4503-4914-7, doi: 10.1145/3038912.3052591.

[6] M. Martens, S. Shen, A. Iosup, and F. Kuipers, "Toxicity detection in multiplayer online games," International Workshop on Network and Systems Support for Games (NetGames 2015) IEEE/ACM, De. 2015, pp. 1–6, ISBN: 978-1-5090-0068-5.

[7] A. Obadimu, E. Mead, M. N. Hussain, and N. Agarwal, "Identifying Toxicity Within YouTube Video Comment," Proc. International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2019) Springer, Jul. 2019, pp. 214-223, ISSN: 0302-9743, ISBN: 978-3-030-21740-2, doi: 10.1007/978-3-030-21741-9.

[8] M. Duggan, "Online Harassment", Pew Research Center. [Online]. Retrieved: August 6, 2019. Available from: http://www.pewinternet.org/2014/10/22/online-harassment/

[9] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," Proc. International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing (PASSAT-SOCIALCOM 2012), IEEE/ASE, Sept. 2012, pp. 71–80, ISBN: 9781467356381 and 978-0-7695-4848-7.

[10] M. N. Hussain, S. Tokdemir, S. Al-khateeb, K.K. Bandeli, and N. Agarwal, "Understanding digital ethnography: socio-computational analysis of trending YouTube videos," Proc. International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2018) Springer, Jul. 2018, Late Breaking Paper. [Online]. Retrieved: August 7, 2019. Available from: http://sbp-brims.org/2018/proceedings/papers/latebreaking_papers/LB_14.pdf

[11] M. N. Hussain, S. Tokdemir, N. Agarwal, and S. Al-Khateeb, "Analyzing disinformation and crowd manipulation tactics on YouTube." Proc. International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2018), IEEE/ACM, Aug. 2018, pp. 1092-1095, doi: 10.1109/ASONAM.2018.8508766.

[12] C. Nobata, J Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," Proc. 25th International Conference on World Wide Web (WWW 2016), ACM, Apr. 2016, pp. 145-153, ISBN: 978-1-4503-4143-1.

[13] Q. Cao, X. Yang, J. Yu, and C. Palow, "Uncovering large groups of active malicious accounts in online social networks," Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS 2014), ACM, Nov. 2014, pp. 477-488, ISBN: 978-1-4503-2957-6.

[14] Y. Chen, Y. Zhou, and S. Zhu, H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," Proc. International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing (PASSAT-SOCIALCOM 2012), IEEE/ASE, Sept. 2012, pp. 71–80, ISBN: 978-0-7695-4848-7, doi: 10.1109/SocialCom-PASSAT.2012.55.

[15] D. Yin et al., "Detection of harassment on web 2.0," Proc. Content Analysis in Web 2.0 (CAW2.0 2009), EPrints, Apr. 2009, pp. 1-7. [Online]. Retrieved: August 7, 2019. Available from: http://www2009.eprints.org/255/6/Yin_etal_CAW2009.pdf

[16] S. O Sood, J. Antin, and E. Churchill, "Using crowdsourcing to improve profanity detection," Proc. Spring Symposium Series of the Association for the Advancement of Artificial Intelligence (AAAI 2012), AAAI Press, Mar. 2012, ISBN 978-1-57735-555-7. [Online]. Retrieved: August 7, 2019. Available from: https://www.aaai.org/ocs/index.php/SSS/SSS12/paper/view/4256/4698

[17] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," Proc. Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017), AAAI Press, May 2017, ISBN 978-1-57735-788-9. [Online]. Retrieved: August 7, 2019. Available from: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665/14843

[18] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving Google's Perspective API Built for Detecting Toxic Comments," arXiv preprint, Feb. 2017, arXiv:1702.08138. [Online]. Retrieved: August 7, 2019. Available from: https://arxiv.org/pdf/1702.08138.pdf

[19] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30-37, 2009.

[20] H. Ma, H. Yang, M. Lyu, and I. King, "Sorec: social recommendation using probabilistic matrix factorization," Proc. 17th ACM conference on Information and knowledge management (CIKM 2008), ACM, Oct. 2008, pp. 931-940, ISBN: 978-1-59593-991-3, doi: 10.1145/1458082.1458205.

[21] W. Peng and B. Xin, "SPMF: A Social Trust and Preference Segmentation-based Matrix Factorization Recommendation Algorithm," arXiv preprint, Mar. 2019, arXiv:1903.04489. [Online]. Retrieved: August 7, 2019. Available from: https://arxiv.org/pdf/1903.04489.pdf

[22] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," Proc. Sixth ACM International Conference on Web Search and Data Mining, ACM, Feb. 2013, pp. 587-596, ISBN: 978-1-4503-1869-3, doi: 0.1145/2433396.2433471.

[23] G. Zhao, M. Lee, W. Hsu, W. Chen, and H. Hu, "Community-based user recommendation in uni-directional social networks," Proc. 22nd ACM International Conference on Information & Knowledge Management, ACM, Oct. 2013, pp. 189-198, ISBN: 978-1-4503-2263-8, 10.1145/2505515.2505533.

[24] M. Jamali and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks." Proc. Fourth ACM Conference on Recommender Systems, ACM, Sep. 2010, pp. 135-142, ISBN: 978-1-60558-906-0, doi: 10.1145/1864708.1864736.

[25] R. Chen, et al., "A Novel Social Recommendation Method Fusing User's Social Status and Homophily Based on Matrix Factorization Techniques," IEEE Access, vol. 7, pp. 18783-18798, 2019, doi: 10.1109/ACCESS.2019.2893024.

[26] M. Ozer, N. Kim, and H. Davulcu, "Community detection in political Twitter networks using Nonnegative Matrix Factorization methods," Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and

Mining (ASONAM 2016), ACM, Aug. 2016, pp. 81-88, ISBN: 978-1-5090-2846-7.

[27] Perspective. [Online]. Retrieved: August 6, 2019. Available from: http://perspectiveapi.com/#/

[28] S. Bay et al., "Responding to Cognitive Security Challenges," NATO STRATCOM COE. [Online]. Retrieved: August 6, 2019. Available from:| https://www.stratcomcoe.org/responding-cognitive-security-challenges.

[29] N. Hug. "Surprise: A Python scikit for recommender systems." [Online]. Retrieved: August 6, 2019. Available from: http://surpriselib.com/

[30] M. Graus, M. Willemsen, and L. Meesters, "Understanding the latent features of matrix factorization algorithms in movie recommender systems," Eindhoven University of Technology, 2011. [Online]. Retrieved: August 7, 2019. Available from: https://pure.tue.nl/ws/files/47007020/712626-1.pdf