Investigating Online Toxicity in Users Interactions with the Mainstream Media Channels on YouTube

Sultan Alshamrani[◇‡], Mohammed Abuhamad[•], Ahmed Abusnaina[◇], and David Mohaisen[◇]

University of Central Florida[◇] Saudi Electronic University[‡] Loyola University Chicago[•]

Abstract

Social media has become an essential platform and source for most mainstream news channels, and many works have been dedicated to analyzing and understanding user experience and engagement with the online news on social media in general, and on YouTube in particular. In this study, we investigate the correlation of different toxic behaviors such as identity hate, and obscenity with different news topics. To do that, we collected a large-scale dataset of approximately 7.3 million comments and more than 10,000 news video captions, utilized deep learning-based techniques to construct an ensemble of classifiers tested on a manually-labeled dataset for label prediction, achieved high accuracy, uncovered a large number of toxic comments on news videos across 15 topics obtained using Latent Dirichlet Allocation (LDA) over the captions of the news videos. Our analysis shows that religion and crime-related news have the highest rate of toxic comments, while economy-related news has the lowest rate. We highlight the necessity of effective tools to address topic-driven toxicity impacting interactions and public discourse on the platform.

1 Introduction

People around the globe adopt social media as an essential part of their daily routine, not only for socializing with each other, but also as a major source of news. Among the different social media platforms, the video-sharing platform "YouTube" has witnessed a massive growth in contents, measured by the number of published videos, as well as their popularity, with a viewership of more than 2 billion monthly users [21]). This massive growth has attracted publishers to deliver their content through video-sharing platforms for a fast delivery of content to viewers, and to enable the social interaction with their viewers, which is enabled by the comment section of videos.

A major feature of video-sharing platforms such as YouTube used for delivering news stories is the interactive experience of the audience. However, users may misuse such a feature by posting toxic comments or spreading hate and racism. To improve the user experience and facilitate positive interactions, numerous efforts have been made to detect inappropriate comments [5]. Despite the efforts focused on detecting inappropriate comments, the associations between various types of toxicity and topics covered in news videos from mainstream media remains an unexplored challenge. This work provides an in-depth analysis of the relationship of such toxic comments and the topics presented on the news. Discovering topics in news videos requires accessing, processing, and modeling the script (*i.e.,* caption) at a fine granularity, to allow the detection of all news topics. Relying on the YouTube categorization feature does not accurately capture the topics of the video. For instance, YouTube has categorized 87.3% of the collected videos as news & politics. To this end, we explored and established topics using the Latent Dirichlet Allocation (LDA) topic-modeling approach that allowed assigning videos to specific topics. Our analysis shows that religion- and violence/crime-related news derive the highest rate of toxic comments constituting 24.8%, and 25.9% of the total comments posted on videos covering these topics, while economy-related news shows the lowest rate of toxic comments with 17.4% of the total comments.

Contribution. This work investigates the online toxicity observed in the comments posted on mainstream

In: A. Editor, B. Coeditor (eds.): Proceedings of International Workshop on Mining Actionable Insights from Social Networks, 20-Oct-2020, published at http://ceur-ws.org

media channels and videos. We summarize our contributions as follows.

- *Data Collection and Ground Truth Annotation:* We collected a large-scale dataset of ≈7.3 million comments posted on more than 14,000 news videos. We manually-annotated approximately six thousand comments to three types of toxicity.
- *Ensemble-based Toxicity Detection:* Using designed and evaluated an ensemble-based approach, that utilizes state-of-the-art techniques for the different stages of our approach incorporating data representation and classification, for detecting various inappropriate comments.
- *LDA-based News Topic Modeling:* Using LDA-based topic modeling, we discovered and defined topics of news videos based on the caption.
- *Topic/Toxicity Association:* Using the discovered topics, we assigned videos to specific topics and explore the topic/toxicity associations for different toxic behaviors. Further, we provided an in-depth analysis of the toxic comments, including their popularity and users' interactions.

2 Related Works

With the growing popularity of online platforms in delivering news [8, 6], the comment section of these platforms has become an important feature where users interact with the contents, contents providers, and each other, to express their opinions on the published contents. The convenience of expressing opinions through the non-restrictive medium of online social platforms may result in misusing such a medium by posting toxic comments [11]. This has led many researchers to investigate different inappropriate behaviors in the comment section of different websites. The majority of the prior research work, however, has focused on designing classification or detection mechanisms for inappropriate comments, while a few have focused on user experience and engagement, as outlined below.

Toxic Comment Classification. Despite various efforts on analyzing toxic contents, identifying distinct behaviors and patterns in this space is a challenge, especially when (1) providing directions for prevention and detection methods, and (2) establishing an association with the comment/content topics. However, there are numerous studies that explored several aspects of toxicity, hate speech, and bias in online social interactions [18, 16, 4, 1].

User Engagement and Interactivity. Another major area in studying user's behavior is using the comments to identify users' engagement with the online news and comments [17, 9, 19]. Diakopoulos *et al.* [3] investigated the relationship between the quality of the comments and both the consumption and production of news on *SacBee.com*, including users motivation for both reading and writing news comments. Ksiazek *et al.* [7] proposed a framework to distinguish between users commenting on contents and those replying to other users to better understand engagement. In this work, and in the same space, we study the correlation between the topic of the news and the type of inappropriate comments, e.g., obscenity and identity hate.

Other noteworthy works that have been conducted on behavioral modeling of YouTube content include [13, 10, 12], although not particularly addressing fine-grained toxicity analysis of mainstream news.

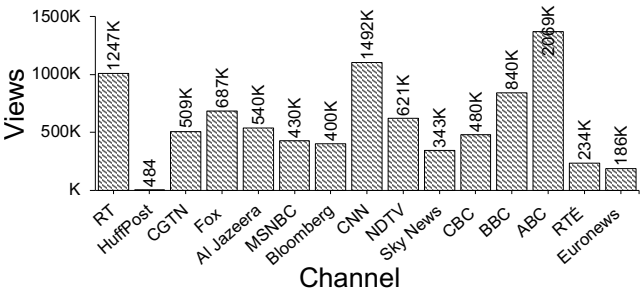


Figure 1: The average number of views per news video for the top-15 mainstream media channels.

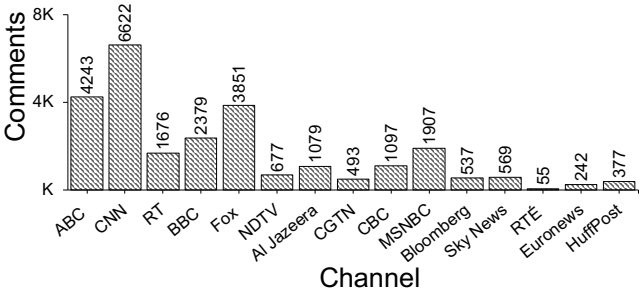


Figure 2: The average number of comments per news video for the top-15 mainstream media channels.

3 Methodology

This section describes the methods used for data collection and representation, toxicity detection, and topic modeling.

3.1 Data Collection and Measurements

The data used in this study consists of comments posted on news videos from YouTube, as well as the captions of these videos. We collected more than 7.3 million comments posted on roughly 14,500 news videos from popular 30 news channels. The collected comments are distributed from early 2007 until October 2019. We were able to extract video captions from

only 10,883 videos, as the remaining videos do not include captions. Moreover, we extended our data collection with the annotated ground truth dataset from the Conversation AI team [2] for the purpose of comment toxicity analysis task.

YouTube News Channels. We collected comments on YouTube videos published by the most viewed mainstream media based on Ranker [14]. We extended our list of mainstream media channels from a Wikipedia list of the most viewed news channels [20]. The final list includes 30 English-speaking news channels from 16 countries.

Data Statistics and Measurements. We collected a total of 7.3 million comments posted by 2,992,273 unique users, and published in the past 13 years (2007 to 2019) where most of the videos were published in 2019, as the trend shows an increase in news video popularity in recent years.

The popularity of the channels used in our study can be seen in the average number of views as shown in Figure 1 for the top-15 most-viewed channels. For instance, videos collected from channels such as ABC, CNN, and RT have a considerably high number of views (*i.e.,* with an average exceeds one million views per video). Intuitively, as the number of views increases, the number of comments is more likely to increase. The average number of comments posted on videos from the most popular mainstream media channels on YouTube is very high as shown in Figure 2. Here, the videos published by CNN, ABC, and Fox news have the highest average number of comments per video which are 6,622, 4,243, and 3,581 respectively. Generally, most of the top-15 channels maintain an average of more than 500 comments per video.

Toxicity-related Annotated Datasets. To study users' behavior in the comment section, we utilized two ground truth datasets to train a machine learning-based ensemble classifier for toxic comment detection and classification: (i) Wikipedia comments created by Conversation AI team [2] and (ii) our own manually-annotated YouTube comments.

- *Wikipedia Ground Truth:* 160,000 comments from Wikipedia Talk pages, manually-annotated by the Conversation AI team, with 143,000 comments labeled as safe, 15,294 toxic, 8,449 obscene, and 1,405 identity hate comments. The labels may overlap, allowing the assignment of more than one label to a toxic comment.
- *YouTube Ground Truth Dataset:* This is an in-house dataset that we created by manually annotating 5,958 random YouTube comments, first into either toxic or safe. The toxic (general class) comments are then mapped to either (*i.e.,* toxic, obscene, or identity hate). The final dataset had 1,832 safe, 4,126 toxic, 2,367 obscene, and 788 identity hate comments.

3.2 Data Preprocessing

For proper data analysis, we initially removed all *non-English contents* across all datasets and eliminated irrelevant characters, tokens, and stop-words. We also removed frequent words appearing in more than 50% of the captions.

3.3 Data Representation

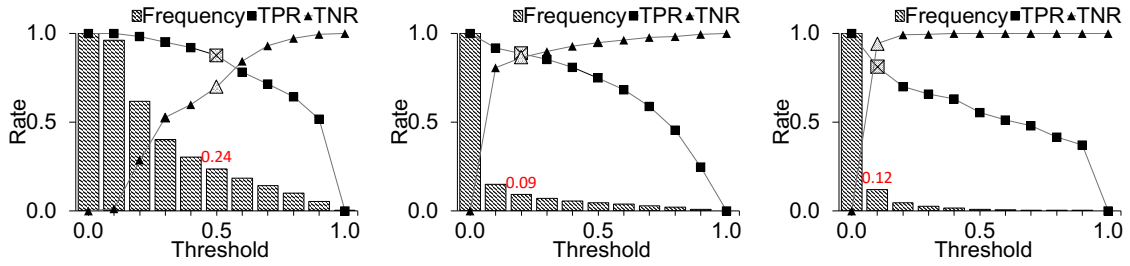
Comments Data Representation. We utilized the pre-trained *Word2Vec* model from Gensim [15]. *Word2Vec* maps words to numerical vectors, and words occurring in a similar context are mapped into similar vectors. Capturing such relationships is possible when acquiring enough data, enabling the *Word2Vec* model to accurately predict the word meaning based on past appearances from the provided context. The comment is then represented as word vectors of size $n \times 300$, where n is the number of words in the comment, with an upper limit of 50 words per comment, as most comments have less than 50 words.

Captions Data Representation. Investigating the topic/comments associations requires defining and understanding the topics raised in videos where the comments are observed. This understanding of topics can be done using topic modeling on captions extracted from videos. For the topic modeling task and topics assignment to videos, we extracted and pre-processed captions from the videos, *i.e.,* transforming captions to lowercase, tokenization, and eliminating irrelevant tokens such as stopwords, punctuation, and words containing less than three characters. After the pre-processing phase, captions are represented using bags of words, in which, words are assigned a unique identifier. To reduce the dimensionality of the bag-of-words, we selected the top 10,000 words to be the caption data representation.

3.4 Toxicity Detection Models

The first task of this study is to detect and classify different toxic behaviors of comments, in order to further investigate their association with the topics covered in the news of which the comments are collected. We inspected comments for three categories of toxicity: *toxic*, *obscene*, and *identity hate*. We utilized a neural network-based ensemble of three models for classifying the three toxic categories.

Deep Neural Network (DNN)-based Architecture. DNN is a supervised learning method that can



(a) Toxic (b) Obscene (c) Identity hate

Figure 3: The evaluation of the ensemble model across categories in terms of TPR and TNR.

discover both linear and non-linear relationships between the input and the output. Comments represented as sequences of word embeddings are fed to the DNN-based models for labeling. The DNN model used in this study consists of (1) an input layer of size (50×300), similar to the shape of the embeddings of the *Word2Vec* representation, (2) two fully connected hidden layers of size 128 with ReLU activation function, and (3) the output layer with one sigmoid.

Dataset Handling and Splitting. Using the two ground truth datasets, we utilized two different approaches to split the datasets for training and evaluating the models. (1) We adopted a 50/50 splitting method for the training and testing of our models using our YouTube ground truth comments datasets. Since the manually-annotated comments dataset is relatively small, the training process is initially done using Wikipedia ground truth comments dataset. Then, each model was fine-tuned using the 50% training dataset of the manually-annotated YouTube comments. (2) We also used 50/50 training/testing splits of the Wikipedia ground truth comments dataset for exploring the effects of different experimental settings. We note that comments can be categorized into multiple toxic categories, *e.g.* one comment can be toxic, obscene, and implies identity hate. Therefore, comments that imply multiple toxic behaviors can be used for training and evaluating multiple models.

3.5 Topic Modeling using LDA

Topic modeling is an unsupervised statistical machine learning technique that processes a set of documents and detects word and phrase patterns across documents to cluster them based on their similarities.

Fine-grained Topics Extraction. We studied the associations between a specific toxic behavior (*e.g.* obscenity) and an extracted topic from videos of mainstream media channels. To do so, we conducted a topic modeling to assign topics to videos based on their caption. This is a challenging task since YouTube categorization is generic and lacks specification of topics covered in the video script. We observed that most

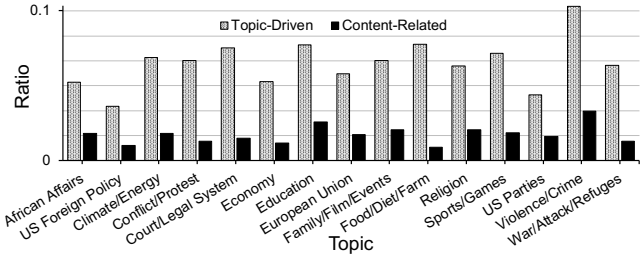


Figure 4: The distribution of the obscene comments over different topics generated by the LDA model.

videos (87.3%) published by the news channels are categorized as *News & Politics*. Based on our analysis of topics appeared in news videos, a variety of topics were captured including war/attack/refugees, violence/crime, sports/games, politics, economy.

LDA Model Settings and Evaluation. The LDA operates using the bag of words representation of caption segments. The topic model receives input vectors of 10,000 bag-of-word representation and assigns topics for each segment. This process includes a training phase that requires setting several parameters such as the number of topics, alpha (the segment-topic density), and beta (topic-word density). To examine the effect of different parameters on the modeling task, we conducted a grid search mechanism to obtain the best configuration of the LDA model that allows for the highest coherence score possible. For the number of topics, we explored the effects of changing the number of targeted topics from 10 to 40 with an increase of 5 topics each iteration. For tuning alpha and beta parameters, we vary the values from 0.01 to 1 with an increment of 0.3 at each step. The LDA-model achieves the best performance using the following settings: $[numberoftopics = 20, alpha = 0.61, beta = 0.31]$ with a coherence score of 0.55.

We manually inspected the frequent keywords of the best-performing LDA output and assigned names and descriptions to them, resulting in various consolidations, and producing 15 distinct topics.

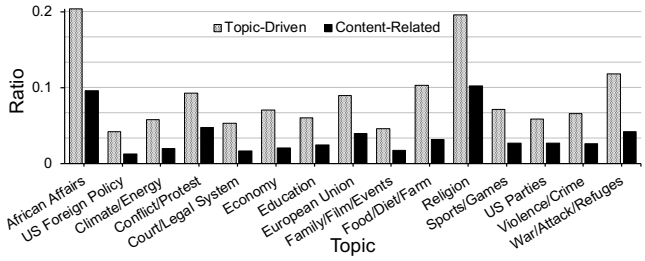


Figure 5: The distribution of identity hate comments over different topics generated by the LDA model.

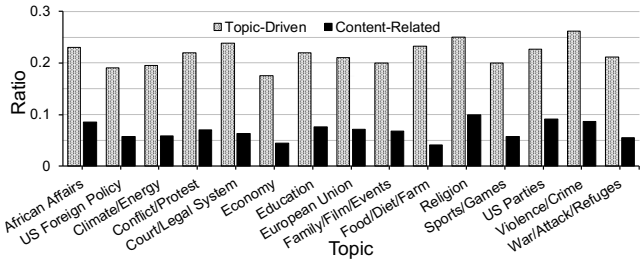


Figure 6: The distribution of the toxic comments over different topics generated by the LDA model.

4 Results and Discussion

4.1 Toxicity Detection and Measurement

❶ **Toxic Comments:** Figure 3(a) shows the performance of the toxic-behavior detection model in terms of TPR and TNR using different classification probability thresholds. We selected the threshold of 0.520 as the best TPR/TNR trade-off with a TPR of 86.2% and a TNR of 71.2%. This model shows that 22.4% of the comments are classified as toxic with a total of 1,648,345 comments.

❷ **Obscene Comments:** The model with a decision threshold of 0.27 achieves a high TPR of 86.6% and TNR of 88.8% for detecting obscene comments. Figure 3(b) shows the results of adopting different thresholds. Applying the model allows the classification of 7.43% of the comments as obscene with a total of 547,222 comments.

❸ **Identity Hate Comments:** Figure 3(c) shows the outstanding performance of the specialized model for detecting identity hate. Using a decision threshold of 0.140, the model achieves a TPR of 74.8% and a TNR of 98.4%. The model shows that 7.03% of the comments are classified as identity hate with a total of 518,213 comments.

4.2 Toxicity and Topics Associations

The detection of toxic behaviors and access to the topic categorization of videos allow us to conduct toxicity/topic analyses. Such associations show whether specific toxicity is topic-driven or derived by other factors. Based on our topic model and ensemble classifier, we examined the presence of toxic, obscene and identity hate comments on each topic of our LDA model.

❶ **Toxic Comments:** Figure 6 shows that the videos discussing topics related to religions or violence/crime have the highest rate of toxic comments, with roughly 25% of the comments are toxic. On the other hand, economy-related news shows the lowest rate of toxic comments with 17% of the total number of comments.

❷ **Obscene Comments:** The violence/crime-related news had the highest number of obscene comments; 10% of the total comments. News covering the United States foreign policy had the least number of obscene comments, with only 3%, as shown in Figure 4.

❸ **Identity Hate Comments:** Among the 15 topics, African affairs and religion news had the highest ratio of identity hate comments; 20% of the comments. While news related to climate/energy and the United States foreign policy have the least number of identity hate comments with about 4% of total comments as shown in Figure 3(c).

Content-related Toxicity. We note that toxic comments can be posted due to several factors and may not be totally driven by the covered topics. In an attempt to relate specific toxic comments with the topics content, we conducted a statistical analysis to measure the commonalities between comments and the content of the caption. For videos of each topic, we obtained the average number of common terms and expressions to be the baseline of indicating the relationship between the topic and the toxic comment. We note that this might not always hold. However, we observed that comments containing a number of common terms with the caption that is higher than the average of common terms in a target topic are more likely to be related to the topics covered in the caption. This analysis produced similar ratios of different toxic behaviors in different news topics.

5 Conclusion

We designed and evaluated an ensemble of models to detect various types of toxicity in comments posted on YouTube mainstream media channels. By analyzing 7 million YouTube comments, posted on 14,506 YouTube news videos, we detected and classified toxic comments with high accuracy, and demonstrated that despite countless efforts in comment moderation taken by YouTube, ≈69% of the collected videos contained toxic comments. We investigated the correlation between the content of news videos and different toxic

behaviors across 15 topics, showing that religion and violence/crime-related news have the highest rate of toxic comments, while economy-related news have the lowest rate of toxic comments. While interesting in its own right from a behavioral standpoint, this study highlights the need for more effective moderation.

Acknowledgement. Work was done while all authors were at the University of Central Florida, and is supported by NRF grant 2016K1A1A2912757 (Global Research Lab). S. Alshamrani was supported by a scholarship from the Saudi Arabian Cultural Mission.

References

[1] Brassard-Gourdeau, É., and Khoury, R. Impact of sentiment detection to recognize toxic and subversive online comments. *CoRR abs/1812.01704* (2018).

[2] ConversationAI. https://conversationai.github.io/, 2019. Accessed: 2019-10-03.

[3] Diakopoulos, N., and Naaman, M. Towards quality discourse in online news comments. In *Proc. of the ACM Conference on Computer Supported Cooperative Work, CSCW* (2011).

[4] D'Sa, A. G., Illina, I., and Fohr, D. Towards non-toxic landscapes: Automatic toxic comment detection using DNN. *CoRR abs/1911.08395* (2019).

[5] Ernst, J., Schmitt, J. B., Rieger, D., Beier, A. K., Vorderer, P., Bente, G., and Roth, H.-J. Hate beneath the counter speech? a qualitative content analysis of user comments on youtube related to counter speech videos. *Journal for Deradicalization*, 10 (2017), 1–49.

[6] GEIGER, A. Key findings about the online news landscape in america. tinyurl.com/y44m63xu, 2019. Accessed: 2020-16-04.

[7] Ksiazek, T. B., Peer, L., and Lessard, K. User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New media & society 18*, 3 (2016), 502–520.

[8] Locklear, M. More people get their news from social media than newspapers. https://tinyurl.com/y8ht3ubr, 2018. Accessed: 2020-16-04.

[9] Ma, Z., Sun, A., Yuan, Q., and Cong, G. Topic-driven reader comments summarization. In *Proc. of 21st ACM International Conference on Information and Knowledge Management, CIKM* (2012).

[10] Mariconti, E., Suarez-Tangil, G., Blackburn, J., Cristofaro, E. D., Kourtellis, N., Leontiadis, I., Serrano, J. L., and Stringhini, G. "you know what to do": Proactive detection of youtube videos targeted by coordinated hate attacks. *Proc. ACM Hum. Comput. Interact. 3*, CSCW (2019), 207:1–207:21.

[11] Massaro, T. M. Equality and freedom of expression: The hate speech dilemma.

[12] Papadamou, K., Papasavva, A., Zannettou, S., Blackburn, J., Kourtellis, N., Leontiadis, I., Stringhini, G., and Sirivianos, M. Disturbed youtube for kids: Characterizing and detecting disturbing content on youtube. *arXiv:1901.07046* (2019).

[13] Papadamou, K., Zannettou, S., Blackburn, J., Cristofaro, E. D., Stringhini, G., and Sirivianos, M. Understanding the incel community on youtube. *CoRR abs/2001.08293* (2020).

[14] Ranker. www.ranker.com, 2019. Accessed: 2019-09-09.

[15] Řehůřek, R., and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proc. of the Workshop on New Challenges for NLP Frameworks* (2010).

[16] Shtovba, S., Shtovba, O., and Petrychko, M. Detection of social network toxic comments with usage of syntactic dependencies in the sentences. In *Proc. of the 2nd International Workshop on Computer Modeling and Intelligent Systems, CMIS* (2019).

[17] Sil, D. K., Sengamedu, S. H., and Bhattacharyya, C. Supervised matching of comments with news article segments. In *Proc. of the 20th ACM Conference on Information and Knowledge Management, CIKM* (2011).

[18] Silva, L. A., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. Analyzing the targets of hate in online social media. In *Proc. of the 10th International Conference on Web and Social Media, ICWSM* (2016).

[19] Tsagkias, M., Weerkamp, W., and de Rijke, M. Predicting the volume of comments on online news stories. In *Proc. of 18th ACM Conference on Information and Knowledge Management, CIKM* (2009).

[20] wikipedia. https://tinyurl.com/y5oyytc8, 2019. Accessed: 2019-09-09.

[21] YouTube. https://tinyurl.com/y9nmv95q, 2020. Accessed: 2020-04-29.