STUDYING USERS INTERACTIONS AND BEHAVIOR IN SOCIAL MEDIA USING NATURAL LANGUAGE PROCESSING

by

SULTAN ALSHAMRANI
M.S. Loyola University Chicago, 2018
B.S. University of Tabuk, Saudi Arabia, 2014

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida,
Orlando, Florida

Fall Term
2021

Major Professor: David Mohaisen

ABSTRACT

Social media platforms have been growing at a rapid pace, attracting users' engagement with the online content due to their convenience facilitated by many useful features. Such platforms provide users with interactive options such as likes, dislikes as well as a way of expressing their opinions in the form of text (i.e., comments). As more people engage in different social media platforms, such platforms will increase in both size and importance. This growth in social media data is becoming a vital new area for scholars and researchers to explore this new form of communication. The huge data from social media has been a massive aid to researchers in the mission of exploring the public's behavior and opinion pursuing different venues in social media research.

In recent years, social media platforms have facilitated the way people communicate and interact with each other. The recent approach in analyzing the human language in social media has been mostly powered by the use of Natural Language Processing (NLP) and deep learning techniques. NLP techniques are some of the most promising methods used in social media analyses, including content categorization, topic discovery and modeling, sentiment analysis. Such powerful methods have boosted the process of understanding human language by enabling researchers to aggregate data relating to certain events addressing several social issues.

The ability of posting comments on these online platforms has allowed some users to post racist and obscene contents, and to spread hate on these platforms. In some cases, this kind of toxic behavior might turn the comment section from a space where users can share their views to a place where hate and profanity are spread. Such issues are observed across various social media platforms and many users are often exposed to these kinds of behaviors which requires comment moderators to spend a lot of time filtering out these inappropriate comments. Moreover, such textual "inappropriate contents" can be targeted towards users irrespective of age, concerning a variety of topics not only controversial, and triggered by various events.

Our work is primarily focused on studying, detecting and analyzing users' exposure to this kind of

toxicity on different social media platforms utilizing state-of-art techniques in both deep learning and natural language processing areas, and facilitated by exclusively collected and curated datasets that address various domains. The different domains, or applications, benefit from a unified and versatile pipeline that could be applied to various scenarios. Those applications we target in this dissertation are: (1) the detection and measurement of kids' exposure to inappropriate comments posted on YouTube videos targeting young users, (2) the association between topics of contents cover by mainstream news media and the toxicity of the comments and interactions by users, (3) the user interaction with, sentiment, and general behavior towards different topics discussed in social media platforms in light of major events (i.e., the outbreak of the COVID-19 pandemic).

Our technical contribution is not limited to only the integration of the various techniques borrowed from the deep learning and natural language processing literature to those new and emerging problem spaces, for socially relevant computing problems, but also in comprehensively studying various approaches to determine their feasibility and relevant to the discussed problems, coupled with insights on the integration, as well as a rich set of conclusions supported with systematic measurements and in-depth analyses towards making the online space safer.

To my family.

ACKNOWLEDGMENTS

This work would not have been possible without the support of many individuals who have enriched my personal and professional experience over the past three years.

I would like to extend my gratitude to my doctoral advisor, David Mohaisen, for his continued support and guidance during the past three years. Coming to the University of Central Florida, I had no prior research experience, and working in the security and analytics lab (SEAL) under David's guidance helped me grow professionally thanks to his guidance, and I am grateful for this opportunity. I would like to also thank my doctoral dissertation committee members, Prof. Yanjie Fu, Prof. Wei Zhang, and Prof. Sung Choi Yoo, for their feedback at the different stages that led to the crystallization of this work: the candidacy, proposal, and final dissertation exam. Their critiques and feedback have been immensely helpful in improving the presentation.

While working at SEAL, I had the opportunity to collaborate with, mentor, and learn from a lot of the current and past members whom I would like to thank (in no particular order): Ahmed, Ashar, Hisham, Mohammad, Rhongho, Saad, Ulku, Necip, Jinchun, Jeman, Samprati, Priyanka, Amin, Anho, Afsah, and Abdurrahman.

Many others in administrative staff and the technical support staff of the department of computer science at UCF have helped me over the past four years. For that, I would like to extend my gratitude to Ernie Gemeinhart for the support before and during the pandemic, and [Nicole Stelter] for helping with the proofing and formatting of this dissertation.

I would like to thank the sponsors of the work reported in this dissertation: Saudi Electronic University in Riyadh, Saudi Arabia, and Saudi Arabian Cultural Mission (SACM) in Washington, D.C., USA. In addition, part of this work was also supported by National Science Foundation (grant number CNS-1809000), National Research Foundation (grant number NRF-2016K1A1A2912757), CyberFlorida (Collaborative Seed Grant), and NVIDIA (GPU Grant).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Understanding the ways humans are interacting and communicating with each other has become more feasible to research thanks to the reliance of people on social media for communication. This has allowed access to the unprompted feelings and opinions of people from different cultures and backgrounds. Such access used to be challenging since researchers used to get such access by collecting users' opinions and feelings through traditional methods such as surveying groups of people. This has provided researchers with a great opportunity of accessing a large scale of users' opinion in people's own words from social media.

Among the various platforms, YouTube is the most popular video-sharing platform and is commonly used by children as an alternative to traditional TV, and as a source of entertainment and educational materials alike. A recent study by [68] reported that 81% of U.S. parents allow their children to use YouTube as an entertainment activity. Another study shows that children under the age of eight spend 65% of their time on the Internet using YouTube [24]. Therefore, researchers have spent enormous efforts understanding the age-appropriate experience of children and adolescents when using YouTube, and have shown that inappropriate contents—such as content with sexual hints, abusive language, graphic nudity, child abuse, horror sounds, and scary scenes—are common, with promoters for such contents targeting this demographic [32, 53, 73]. Parents and custodians trust children-oriented YouTube channels, such as Nick Jr., Disney Jr., and PBS Kids, to present educational and entertaining material for their children even with no supervision. Even when watching videos from trusted family-friendly channels, the written contents, such as user comments, might contain inappropriate language or malicious URLs that could harm or affect the children's behavior. The limited work on the textual contents of YouTube videos targeting kids, as opposed to the various efforts on understanding YouTube's video/audio contents, creates the need for comments-based studies.

Moreover, the popularity of this platform has attracted news publishers to deliver their content

through video-sharing platforms for a fast delivery of contents to viewers, and to enable the social component of interaction with their viewers, which is enabled by the comment section of videos. A recent study by Smith *et al.* [68] shows that the number of users getting their news from YouTube, for example, has nearly doubled between 2013 (20%) and 2018 (38%), and 53% of YouTube users consider YouTube as an important source for understanding current events from around the world. A major feature of video-sharing platforms such as YouTube used for delivering news stories, as compared to the traditional medium, 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 [23, 32]. Despite these efforts focused on detecting inappropriate comments, the associations between various types of toxicity and topics covered in news videos from mainstream media remain an unexplored challenge. Such association could greatly benefit news publishers and moderators to focus on certain events to efficiently be able to eliminate any hate or racism before it spreads on the platform.

In the effort of understanding users' behavior at this time, one of the events that have changed people's life and behavior, including their interaction and communication is the outbreak of COVID-19. With social distancing and other measures being enforced in an attempt to limit the spread of COVID-19, social media has become a medium for communication, expression, and entertainment. In such a period, data science and mining play a central role in understanding the effect of the pandemic on users, and their behavior and perception. Understanding the trends, and people's perceptions, using sentiment analysis allows a better understanding of users' behavior, and their reaction toward a particular topic is a marvelous area to explore.

Motivation

The focus of this work is to conduct a comprehensive study and analysis of user's behaviour and interaction on social media to tackle some trending social issues using natural language processing

techniques. In particular, we have collected a large dataset from YouTube comments to measure and analyze kids exposure to malicious and inappropriate contents. We have also studied the correlation of different types of streamed news on YouTube with different toxic behaviors given insights on what kind of news that people are more likely to post toxic comments on. Moreover, and in the current outbreak of Covid-19, we have studied the public behavior on Twitter by studying the topics raised during the outbreak and the users' sentiment and perception towards these topics during the pandemic. Here is a summary of the social issues that we addressed in this work:

Kids exposure to inappropriate and malicious comments on YouTube. Since children are using YouTube as an alternative to traditional TV, this makes them more susceptible to encountering hate, racism, and obscenity in the comment section. They might also be tricked into clicking on malicious URLs embedded in the comments. Therefore, and in order to increase parents and custodian awareness we have collected a large dataset of comments posted on kids' videos. Using this dataset, we investigated the presence of both inappropriate languages as well as the presence of malicious URLs.

Toxicity in Users Interactions with the Mainstream Media Channels on YouTube. In order to understand and ease the process of moderating comments on YouTube, we collected a massive dataset of YouTube news scripts and the comments posted on them. The purpose of such data is to draw the association of different types of toxicity such as hate, obscenity with different news topics such as politics, economy, etc.

Users Engagement on Twitter During the COVID-19 Pandemic. Through the outbreak of Covid-19 people are heavily using different social media platforms to socialize after certain measures took place to limit the spread of the disease. To study such change in the users' perception towards different topics we have collected all tweets from 14 major cities in English and investigate the topics they discuss before and after the outbreak and how it has changed during the pandemic.

Approaches

Toxicity Classification Models. In detecting toxicity on the YouTube comments, our approach adopts an ensemble of classifiers to predict different toxic categories using DNN models. Based on our experiment, DNN performs very well in terms of identifying different toxic categories as opposed to CNN and RNN. We found that different pre-trained models for feature representation, such as Glove and Gensim, work better in certain scenarios for identifying certain age-inappropriate comments categories (i.e., Glove with DNN for identifying threat comments). The ensemble uses DNN for identifying five age-inappropriate categories, DNN with gensim Word2Vec for identifying toxic, obscene, insult, and identity hate categories, and DNN with Glove Word2Vec for identifying threat category.

Topic Modeling Using Latent Dirichlet allocation LDA. For both tasks of extracting topics from news transcript and the tweets posted during the pandemic, we employed LDA using the bag of words representation. The topic model receives input vectors of 10,000 bag-of-word representation and assigns topics for each piece of text. 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.

Sentiment Analysis Using BERT. In essence, Bidirectional Encoder Representations from Transformers (BERT) [16] is a language model that benefits from the attention mechanism used in the transformer architecture [80]. This attention mechanism has two six-layers of encoders that have the ability to learn contextual relation between words in a given text, as well as six layers of decoders that generate the needed output for a given task. Since BERT is a language model, it uses only the encoder part of the transformer. By adding a new layer to the core model. In our implementation of BERT, we used the same structure and model configuration of the original work. For

more details, we refer the reader to the original research paper in [16]. In our study, tweets are separated per sentence, with two special tokens indicating the start of the tweet and the end of each sentence. Then, each tweet is fed into the trained BERT model, providing the embedding of each word in the tweet considering its surroundings. The output of the BERT model is a one-neuron output layer with a sigmoid activation function for binary classification signaling the polarity of the tweet, *i.e.* either positive or negative sentiment.

Organization. This thesis is organized as follows: We visit the literature and outline the notable related works in chapter 2. In chapter 3, we study the exposure of children and adolescents to inappropriate and malicious comments on YouTube. In chapter 4, we investigated Online toxicity in users' interactions with the mainstream media channels. In chapter 5, we explored and analyze users' behavior before and during COVID-19 using trending topics and sentiments.

CHAPTER 2: RELATED WORK

In the following, we discuss the works relevant to our work. We started by discussing studies focus on children safety on YouTube covered in the literature, followed by several works related to the usage of YouTube as an alternative to traditional TV channels. Then, we presents the latest works on studying and analyzing toxicity on social media comments. Finally, we show several studies that focus on users' behavior and and engagement with different online platforms, followed by works that study such behavior in the light of COVID-19.

Children Safety on Social Media

Recently, several studies have been conducted with the aim of exploring the effects of social media on children, since the use of social media has become a significant part of their daily routines. To ensure the safety of kids on YouTube, Alshamrani *et al.* [5] studied the exposure of children to malicious URLs on videos targeting young users. Another study by [49] which has encouraged parents to understand and be aware of the various possible offline and online behaviors of their children, such as cyber-bullying, privacy issues, sexting, and Internet addiction.

Among other social media platforms, YouTube has been the subject of many studies since it is considered the most popular social media platform in the United States [67], and the second-largest search engine after Google worldwide [46]. Studying the appropriateness of contents being presented to children on YouTube was first considered, to the best of our knowledge by Kaushal *et al.* [32] who studied kids-unsafe contents and promoters. The authors provided a framework for detecting unsafe contents using measures calculated on the video, user, and comment levels with an accuracy of 85.7%.

Another work by Papadamou *et al.* [53] shows that inappropriate toddler-oriented videos are common and likely to be suggested by YouTube's recommendation system. Using manually-annotated videos, the authors investigated the detection of inappropriate content (containing sexual hints,

abusive language, graphic nudity, child abuse, horror sounds, and scary scenes) collected from videos targeting kids using deep learning algorithms to achieve an accuracy of 84.3% for this task. More recently, Tahir *et al.* [73] demonstrated that even children-focused apps, such as *YouTube Kids* which is considered a kids-safe platform, are prone to compromise with inappropriate videos.

Social Media as an Alternative to News Outlets

With the increased popularity of online platforms in delivering news to users [38, 27], the comment section of these platforms has become an important place where people interact with the contents, contents providers and each other in order to express their opinions and thoughts. The convenience and freedom of expressing opinions through the non-restrictive medium of online social platforms may result in misusing such a medium by posting toxic comments [41].

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 been mainly focused on designing classification or detection mechanisms for inappropriate comments, while a few have focused on the user experience and engagement with different social media platforms. In the following, we reviewed these works.

Toxic Comment Classification

Despite the several efforts on analyzing and understand toxic contents, identifying distinct behaviors and patterns is a challenging task, especially when (1) providing directions for prevention and detection methods, and (2) establishing an association with the topics in which the comments appear. However, there are numerous studies that explored several aspects related to toxicity, hate speech, and biases in online social interactions [66, 64, 21, 12, 48]. To study inappropriate comments online, Ernst *et al.* [23] applied a qualitative content analysis on randomly selected comments posted on videos about "Concepts of Islam". The results showed that most of the comments

dealt with prejudices and stereotypes towards Muslims.

Another form of toxic comments is the use of profanity or obscene language. For instance, Sood *et al.* [70] built a system that outperforms the performance of list-based profanity detection techniques by employing crowdsourcing using Amazon Mechanical Turk. While several researchers proposed different techniques and detection mechanisms to detect hate speech, others tried to enhance or challenge the current approaches [18, 79].

User Engagement and Interactivity

Another major area in studying user's behavior is using the comments to identify users engagement and interaction with the online news and comments [65, 39, 78]. Diakopoulos *et al.* [17] investigated the relationship between the quality of the comments and both the consumption and production of the news on *SacBee.com*.

Moreover, they investigated users motivation for both reading and writing news comments. Similarly, Ziegele *et al.* [89] conducted a qualitative interview with users posting comments on news articles in order to understand their motivation for being interactive in the news comment section. The study analyzed 1,580 comments and found that uncertainty, controversy, comprehensibility, negativity, and personalization are the most important factors for the interactivity in the news comment section.

However, in some cases, the content of the news is not the main reason for the interactions in the comment section, whereby the user-to-user response can be the main driver of interaction. In this regard, Ksiazek *et al.* [35] proposed a theoretical framework to distinguish between user–content—user commenting on content—and user-to-user—user replying to another user's comment—modes of interaction to better understand users engagements. In this work, we studied the correlation between the topics raised in news videos and different types of inappropriate comments, such as insult, obscenity, identity hate, etc., which is an additional gap we fill in this space.

User Engagement During the Pandemic

The significant impact of COVID-19 on many aspects of our daily lives has encouraged many researchers to investigate users' perceptions and behavioral coping activities during the pandemic. This interest in studying users' perceptions is motivated by the need for providing well-informed measures to keep people safe and address the spread of misinformation about the pandemic. Such misinformation affects peoples' lives and the way they behave. This, in turn, leads to unsafe practices that result in an increased COVID-19 case, developing further physical or mental health issues.

Studying the spread of COVID-19 misinformation on Twitter was presented by Kouzy *et al.* [34]. The authors studied the amount of COVID-19 misinformation on Twitter by analyzing tweets from hashtags, terms related to COVID-19. The study employed statistical analysis, comparing terms and hashtags to identify certain tweets and account characteristics. Using a dataset of 673 tweets, the authors reported 153 tweets to have some degree of misinformation, and 107 to be posted from unverifiable sources.

Understanding and monitoring people concerns and needs can help authority health officials to instantiate better guidance and measures to contain the spread of the virus. Therefore, in this study, we provide in-depth analysis and insights into how the pandemic has impacted people's behavior towards topics discussed on Twitter. Moreover, we analyze trends and shifts of emotions and feelings observed when discussing these topics before and after the pandemic. To do so, we utilize state-of-the-art techniques to detect, model, and track different topics from tweets collected in time-frame covering periods before and after the COVID-19 outbreak. After observing the major topics discussed during the data collection period, we adopted a deep learning-based sentiment analysis to study the people's perceptions of these topics before and after the pandemic

CHAPTER 3: MEASURING THE EXPOSURE OF CHILDREN TO MALICIOUS AND INAPPROPRIATE COMMENTS

The influence of social media on the intellectual and emotional well-being of children and adolescents has been the focus of many studies in recent years, with social media being a central daily activity in children and adolescents' lives alike [26]. Among the various platforms, YouTube is the most popular video-sharing platform and is commonly used by children as an alternative to traditional TV, and as a source of entertainment and educational materials alike. A recent study by [68] reported that 81% of U.S. parents allow their children to use YouTube as an entertainment activity. Moreover, another study shows that children under the age of eight spend 65% of their time on the Internet using YouTube [24]. Therefore, researchers have spent enormous efforts understanding the age-appropriate experience of children and adolescents when using YouTube, and have shown that inappropriate contents—such as content with sexual hints, abusive language, graphic nudity, child abuse, horror sounds, and scary scenes—are common, with promoters for such content targeting this demographic [32, 53, 73]. Parents and custodians trust children-oriented YouTube channels, such as Nick Jr., Disney Jr., and PBS Kids, to present educational and entertaining material for their children even with no supervision. However, children can be exposed to inappropriate and disturbing videos, suggested by the YouTube recommendation system, as children are tricked to click on innocent-looking thumbnail [53]. To ensure their well-being and safety, it is important to study the exposure of children and adolescents to inappropriate material presented on YouTube, including visual, audio, and written content. Even when watching videos from trusted family-friendly channels, the written contents, such as user comments, might contain inappropriate language that could influence the children's offline behavior. The limited work on YouTube textual contents, as

The work in this chapter has been published in the 8th International Workshop on Natural Language Processing for Social Media (SocialNLP 2021); held in conjunction with WWW 2021, and the Third ACM/IEEE Workshop on Hot Topics on Web of Things (HotWoT 2020); held in conjunction with ACM/IEEE SEC 2020.

opposed to the various efforts on understanding YouTube's video/audio contents, creates the need for comments-based studies.

Our study explores measuring the exposure of children and adolescents to age-inappropriate comments posted on videos of the top-200 children shows [57]. This task is challenging for several reasons. First, studying comments on children's videos requires manually collecting channels and shows targeting this demographic, knowing YouTube categories are not established by age-group but rather by the topic they present. Second, assigning age groups to the collected videos can be daunting in measuring exposure by separate groups. Third, the lack of a ground truth dataset for safe and inappropriate content posted on such videos makes it difficult to allow appropriate machine-learning models to capture the children's exposure on a large scale. Considering the variety of age-inappropriate content for children, building a unified system for detecting such content is challenging.

To address those challenges, we built a large collection of YouTube comments on children-oriented videos for the top 200 shows categorized by different age groups [13]. We extended the dataset with ground truth data from different sources to establish five age -inappropriate categories; toxic, obscene, insult, and identity hate. The used ground truth dataset compresses annotated data provided by Conversation AI on Wikipedia's comments, and our own manually-annotated data from YouTube comments posted on children videos. We leveraged natural language processing and machine learning techniques to construct an ensemble of models, each of which specializes in detecting a specific inappropriate category. The models are trained and tested on ground truth samples, and separately and collectively achieve remarkable results. Utilizing our ensemble, we uncovered a large number of age-inappropriate comments among those posted on children YouTube videos. Measuring the exposure by age group, our results show that children between 13 and 17 years old are the most exposed to such content. For inappropriate categories, toxic-related comments are the most common, with 15.54% out of the total comments, then insult (7.96%) and obscene (6.84%).

Contribution. This work contributes to measuring the exposure of children and adolescents to

inappropriate content present in the kids' YouTube videos comments. We summarize our contribution as follows:

- We collected a large-scale dataset of comments on children's YouTube videos from the top-200 ranked children shows. The list of shows, retrieved search results, categorization of shows by age group, and other artifacts related to the data collection process are manually vetted.
- We built a manually-annotated ground truth dataset collected from comments posted on children's videos, which includes about 6,000 comments.
- Leveraging natural language processing and deep learning techniques, we designed and implemented an ensemble of classifiers to detect five age-inappropriate contents. Models of the ensemble are trained, fine-tuned, and evaluated using the ground truth dataset.
- Adopting the ensemble classifier on the YouTube comments domain, we detected and measured children's exposure to inappropriate comments.
- We provided an in-depth analysis of children's exposure to inappropriate content in terms of age groups, user interactions, and YouTube video channels.

Dataset

Our dataset includes YouTube comments and two datasets of ground truth, one from the Conversation AI team and another one annotated by our team for ground truth from the YouTube comments. For the YouTube comments, we collected more than 3.7 million comments posted on roughly 10,000 children's videos, distributed over the period from January 2005 until March 2019.

Children's Shows. We collected comments on videos of the top-200 children's shows based on Ranker [58], a crowdsourced platform that relies on millions of users to rank a variety of media contents such as shows and films. The list of shows was originally made by Ranker TV and

received more than 1.2M votes, and has 380 kids' shows. Among them, we selected the top 200 shows. We augmented our list with part of Wikipedia's list of cartoon shows.

Collection Approach. Using YouTube APIs, we extracted the top-50 videos of the search results on every show on our list. Using each retrieved video's ID, we also used the API to obtain video statistics, such as the number of views, likes, dislikes, etc. We used YouTube Comments API to collect all comments from the videos. In total, we collected more than 3.7 million comments from 10,000 videos.

Age-Appropriateness of Children's Shows. We defined age appropriateness as the adequate age group to be the subject of the show. Defining the age appropriateness for children's shows is challenging, since most shows do not specify the target age group. Therefore, we used *Common Sense Media* [13], a non-profit organization that provides education and advocacy to families on providing safe media for children, as the main source for defining the age group of the targeted children's shows. Using *Common Sense Media*, we were able to retrieve the appropriate age group for most of the kids' shows on our list. However, a few shows do not appear in *Common Sense Media*, and for those we turned to IMDB [29], an online database of information about different types of media such as films, television programs, home videos, video games, etc., to obtain the age group for those particular shows. Some shows have different versions, each for a certain age group, therefore the age group is assigned based on the most prevalent version in the YouTube search. Some other kids shows are assigned an age group based on their respective categories, e.g., Loony Tunes (a well-known collection of cartoons for age 7+). We note that we conducted a manual inspection on the age appropriateness for the retrieved top-50 results on each show to define non-kids contents and assigned them to 17+ age group, which is the highest age group in our dataset.

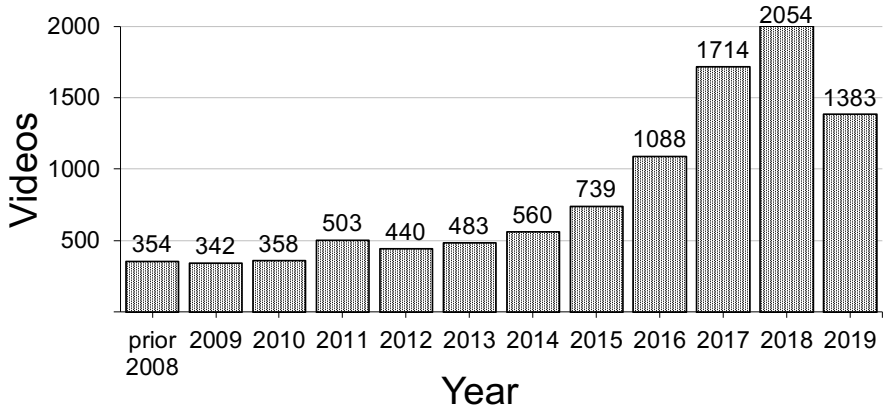


Figure 3.1: The publish date distribution of the collected YouTube kids' videos.

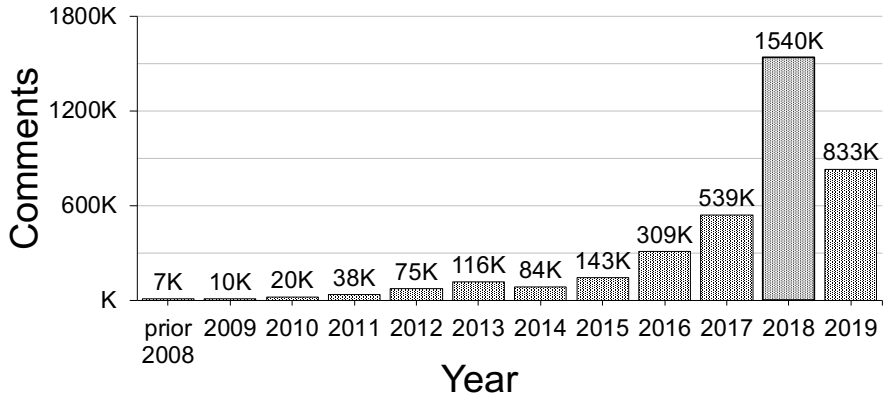


Figure 3.2: The distribution of YouTube kids' videos comments over past years.

Data Statistics and Measurements

Here we provide general statistics of our data. The collected YouTube comments were posted by more than 2.5 million users on about 10,000 videos from more than 3,000 different channels. These retrieved videos have an average viewers count of roughly 2.4 million views and an average comments count of 8,068 comments per video. Observing the publishing date of the videos in our collection, Figure 3.1 demonstrates the rapid increase in children's videos over the past few years. The figure shows an increase in popularity of five folds in ten years from 2008 (with 354 videos) to 2018 (with 2,054 videos). This rapid growth in popularity is observed through the first three

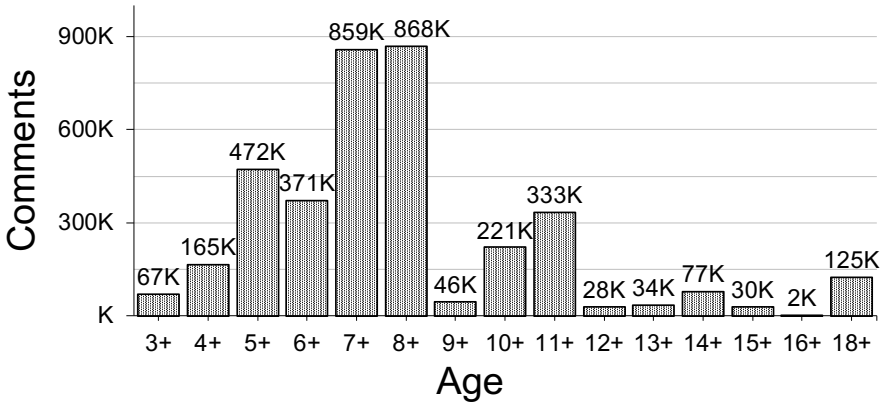


Figure 3.3: The distribution of YouTube kids' comments over different ages.

months of 2019 with 1,383 videos included in our collection (by March of 2019). We note that the collection of YouTube videos is based on their relevance, and not the publishing date nor the view count; this is also the case when retrieving videos from the top-50 search result and when querying the targeted shows. The search results do not always reflect the popularity. However, the top-ranked videos are often characterized by bursts of popularity [25]. Generally, a consistent trend is observed in the year-over-year increasing number of videos included in our collection. Similar patterns are observed with the number of comments from around 7,000 comments on videos prior to 2008 to more than 1.5 million comments on videos from 2018. This growth is steady through the first three months of 2019 as illustrated in Fig. 3.2. We also provided the distribution of comments across the age groups as shown in Fig. 3.3 where most of the collected comments were posted on videos for kids between the age of five and eight (a total of approximately 2.5 million comments).

Age-Appropriateness of Contents. Contents that are regarded as age-appropriate for children and adolescents ideally should not contain toxic words or imply an insult, threat, identity hate, or obscenity. To study the appropriateness of YouTube comments, we collected ground truth datasets to establish a baseline for modeling contents with different labels (i.e., toxic, obscene, insult, threat, and identity hate). The ground truth data includes: (1) labeled comments from Wikipedia that is manually annotated by Conversation AI, a research team started by Jigsaw and Google to provide tools and solutions for improving online conversions; (2) labeled comments posted on YouTube

Table 3.1: The distribution of the dataset. The comments are from two sources: Wikipedia and YouTube. 5,940 YouTube comments are manually labeled for the evaluation of the models.

Source	Dataset	Count
Wikipedia	Safe Comments	143,000
	Toxic	15,294
	Obscene	8,449
	Insult	7,877
	Threat	478
	Identity hate	1,405
YouTube	Unlabeled	$\approx$ 3,700,000
	Safe Comments	1,832
	Toxic	4,126
	Obscene	2,367
	Insult	1,650
	Threat	550
	Identity hate	788

videos targeting children that are manually annotated for the purpose of this study.

(1) Wikipedia Ground Truth Toxic Dataset. We used the manually-annotated dataset provided by Conversation AI, with approximately 160,000 comments from Wikipedia Talk pages of which approximately 143,000 comments are labeled as safe, while the remaining are labeled to have different types of toxicity (i.e., 15,294 toxic, 8,449 obscene, 478 threat, 7,877 insult, and 1,405 identity hate). A summary of the collected data is provided in Table 3.1.

(2) Manually Annotated Ground Truth. We manually annotated 5,958 YouTube comments posted on YouTube videos for the evaluation of the ensemble. The total number of the manually labeled comments is distributed as follows: safe: 1,832, toxic: 4,126, obscene: 2,367, insult: 1,650, threat: 550, and identity hate: 788.

For our manual labeling, we used several explicit rules. Each comment was labeled as either toxic or safe. A toxic comment may belong to one or more of unsafe category; obscene, threat, insult, or identity hate. A comment is considered obscene when it is morally offensive in a sexual way, or when it has socially offensive words. When such an offensive language is used against

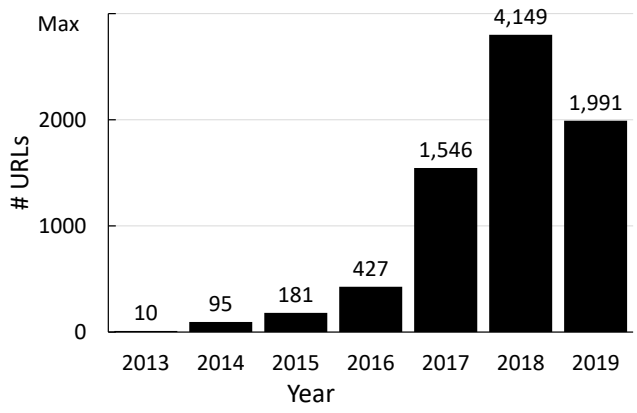


Figure 3.4: The distribution of the collected URLs over the years.

or to describe other users, video publishers, or anyone else, the comment is considered as an insult. When such an offensive language is directed to another group of people, by imposing a negative stereotype or prejudices about people based on their race, color, or ethnicity, the comment is considered as an identity hate.

The annotation is challenging since identifying identity hate is highly subjective [50] [61]. Some comments did not have any profanity or offensive language, but implied a threat to other users or the video publisher; we labeled such a comment to be a threat. In the annotation process, we encountered comments that are socially unacceptable and are age-inappropriate, however, they do not belong to any of the four unsafe categories, and so we labeled them as *toxic* only. The manual labeling has been done by the same annotator (lead author of this work), upon refining the above ruleset. We avoided using multiple annotators across different folds of the manually-labeled dataset, and rather pursued this slow labeling method, to avoid inconsistency and subjectivity in interpretation against the predetermined labeling rules.

Ground Truth: Safe Dataset. For safe content, we used our labeled safe YouTube comments as well as safe-labeled comments from the Conversation AI team dataset, which include roughly 143,000 comments in total.

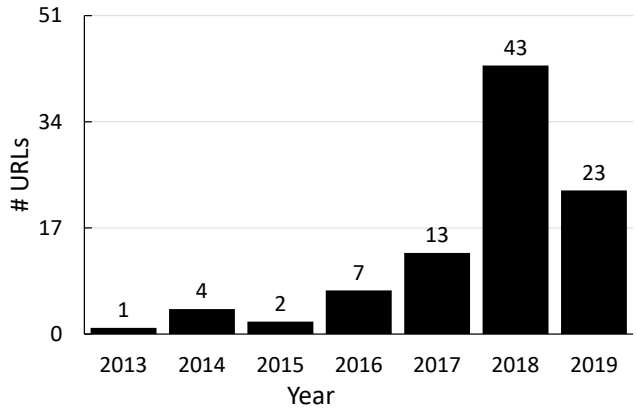


Figure 3.5: The distribution of the inappropriate URLs over the years.

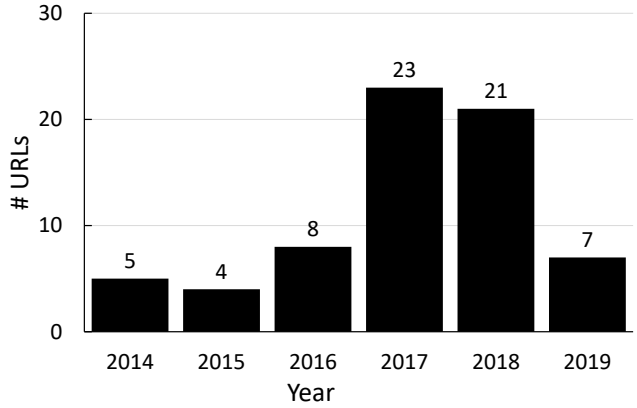


Figure 3.6: The distribution of the Malicious URLs over the years.

URLs Extraction

In order to make sure that the URLs embedded the comments section are not malicious nor do they promote inappropriate content such as adult websites, we investigated the URLs on children's videos and their potential risks on children. We used a regular expression to extract possible URLs within the comments. In the collected dataset, we extracted 8,677 URLs, associated with 1,628 videos. Figure 3.4 shows the number of URLs extracted per year. Notice that there is an increasing trend of the number of URLs embedded in the comments, shedding the light on the importance of monitoring the content of the comment, particularly in children-oriented channels

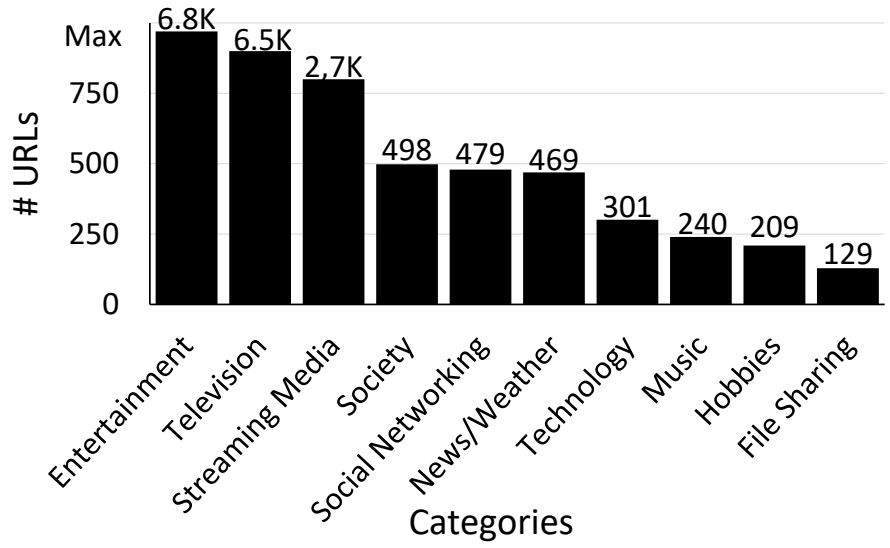


Figure 3.7: The top 10 IAB Categories associated with the collected URLs.

URL Topic Categorization. We extracted the topics associated with the embedded URLs to understand their effects on children's exposure to various contents. In particular, we used Webshrinker [15], a machine learning-powered domain data, and threat classifier, to obtain the Interactive Advertising Bureau (IAB) categorization of the domains of the URLs. To this end, we extracted 107 different categories associated with the URLs. Figure 3.7 shows the top ten categories associated with the URL. Note that a URL may be associated with one or more category, based on IAB categorization. Note also that entertainment, television, alongside streaming media, were the most common categories within the URLs.

Malicious URL Extraction. In the context of videos targeting kids, the chance that the audience will blindly click on the URLs posted on videos is very high. Such behavior may allow attackers to gain information or access to private resources on the victim's device. Therefore, we extracted all the URLs within the collected comments and checked whether the given URL is valid or not by accessing the website and checking the response of the HTML request. If the returned response status code is 200 (success), we then forward the URL to VirusTotal API [82] to check whether it is benign or malicious as well as the URL's associated attributes. Those attributes include "malicious",

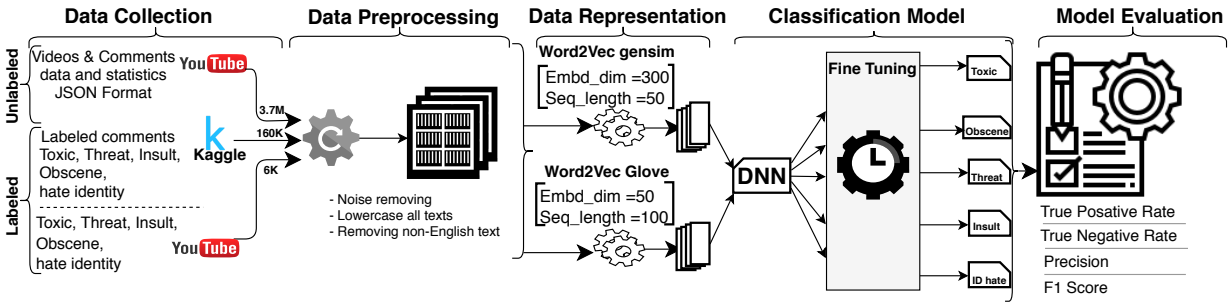


Figure 3.8: The ensemble pipeline. The system design consists of five stages, starting from data collection and labeling, followed by the preprocessing of the data to generate efficient representation. Then, ensemble of five classification models are used for comments classification. Further, the models are evaluated using four evaluation metrics.

"malware", and "phishing". Malicious websites contain exploits or other malicious artifacts, malware websites are used for malware distribution, and phishing websites are used for stealing users' credentials or private information.

Data Preprocessing

Several preprocessing steps are taken before the final data representation, modeling, and evaluation, and to ensure clean and proper representation of the collected data. YouTube comments are the focus of this study, which we addressed with the preprocessing steps as follows: (1) We initially removed all *non-English contents* across all datasets, and limit our analysis to English comments. (2) We eliminated unwanted characters and tokens, e.g., punctuation, and other characters that represent or encode emojis. The entire pipeline employed the system can be sen in Figure 3.8.

Comments Data Representation. In order to perform an analysis of textual data, we first transformed this data into an embedding (i.e., numerical representation) that can be used by machine learning models. Such a representation allows the machine learning models to learn and capture different patterns of the text. We utilized different data representation methods, namely, *Word2Vec* [44] and *Glove* [54].

Pre-trained Word2Vec. Using the pre-trained model for comments representation, we have the following two cases of distinct models. **(1) Gensim:** This technique transforms textual data by examining word statistical co-occurrence patterns within a corpus of the provided textual documents. Examining different configurations for both word embedding and the document vector. We found that the highest accuracy can be achieved using a size of 300 for the word embedding and the document vector size of 50. **(2) Glove:** This technique is an unsupervised learning algorithm used to generate numerical vector representations for words. The training process is done on aggregated global word-word co-occurrence statistics from a corpus. We used Glove to represent the comments; similar to Gensim, we tried different configurations and selected the configuration with the highest accuracy, using a size of 50 for the word embedding and 100 for the document vector.

Ensemble Classification Models

To understand and measure children's exposure to inappropriate comments on YouTube videos by first identifying them, we adopted an ensemble classifier to build five specialized models for classifying five unsafe categories: *toxic*, *obscene*, *threat*, *insult*, and *identity hate*. The models are trained, in a supervised manner, using the Wikipedia toxic comments dataset and the manually annotated ground truth of YouTube comments. Each model predicts whether an input belongs to a specific category, functioning as a binary classification task. We note that a comment can belong to one or more categories (e.g., toxic, insult, and identity hate simultaneously), thus the output of the ensemble is positive if the comment is labeled as at least one age-inappropriate category.

Our approach adopts an ensemble of classifiers to predict different age-inappropriate categories using DNN models. Based on our experiment, DNN performs very well in terms of identifying different age-inappropriate categories as opposed to CNN and RNN. We found that different pre-trained models for feature representation, such as Glove and Gensim, work better in certain scenarios for identifying certain age-inappropriate comments categories (i.e., Glove with DNN for identifying threat comments). The ensemble uses DNN for identifying five age-inappropriate

Table 3.2: The performance of the ensemble model on Wikipedia comments, and the fine-tuned models across different metrics.

Class	Wikipedia			Fine Tuned		
	Recall	Prec	F1	Recall	Prec	F1
Toxic	92.5	82.5	87.2	93.5	83.1	88.0
Obscene	81.9	82.9	82.4	86.6	83.5	85.0
Threat	64.4	43.7	52.1	71.3	42.3	53.1
Insult	74.5	55.7	63.3	66.7	64.4	65.6
Identity hate	53.9	89.8	67.4	74.8	87.8	80.8
Overall	73.4	70.9	70.4	78.5	72.2	74.5

categories, DNN with gensim Word2Vec for identifying toxic, obscene, insult, and identity hate categories, and DNN with Glove Word2Vec for identifying threat category.

We feed Word2Vec vectors of the comments to the first input layer in the network while the output layer has a single node for binary classification to predict whether the provided comment belongs to a certain class or not. Our model architecture is composed of two dense layers of size 128 units with a ReLU activation function, each followed by a dropout operation with a rate of 20%. The last layer is fully connected to a sigmoid function, which generates real values in the range (0,1) using the function $sigmoid(z) = 1/(1 + e^{-z})$. Since the output $\{y \in \mathbb{R} \mid 0 \leq y \leq 1\}$, determines the probability of assigning an input to the target, a threshold can be defined for target $\bar{y}$ assignment (e.g., a commonly-used threshold is 0.5 where $\bar{y} = 1\ if\ y \geq 0.5$). We explored different thresholds for each category to optimize the true negative rate and the true positive rate.

Model Training Settings

We used the entire Wikipedia annotated comments to train five models, each of which is specialized in detecting one age-inappropriate category. Then we fine-tuned the trained model using 50% of our manually labeled comments from YouTube, by only retraining the last layer of the model. We then used the other half for the evaluation of the models. The training process is guided by

minimizing the binary-cross-entropy as follows:

$$\text{loss}(\theta) = \frac{-1}{N}\sum_{i=1}^{N}[y_i \times \log(p_i) + (1 - y_i) \times \log(1 - p_i)],$$

where p_i is the conditional probability $p(y_i|x_i, \theta)$ for a target y_i given an input x_i and a set of parameters θ, i is the i-th record, and N is the total number of records in the training set. The optimization is done using ***RMSprop*** optimizer, a stochastic optimization algorithm, with a learning rate of 10^{-3} without decaying over time. We used a mini-batch approach with a batch size of 128, and for preventing the overfitting we used dropout regularization with a dropout rate of 0.2. The termination criterion is set to be a specified number of training iterations, which is set to 100 for all models.

Evaluation Metrics. This study uses four evaluation metrics, which are *Precision*, *F1-score*, *True Positive Rate* (TPR), and *True Negative Rate* (TNR). Precision represents the percentage of which a model was correct in predicting the positive class (P = TP/TP+FP). F1-score is the harmonic mean of the precision and recall, and is expressed as (F1-score $=$ 2TP/2TP+FP+FN) where TP, FP, and FN represent True Positive, False Positive, and False Negative, respectively. The TPR is the proportion of the positive predictions, positive labeled-data correctly predicted to be positive, form the total positive-labeled data (TPR $=$ TP/TP+FN). The TNR is the proportion of the negative predictions, negative labeled-data correctly predicted as negative, from the total of negative-labeled data (TNR $=$ TN/TN+FP).

Results and Discussion

In this section, we review the results of the ensemble for classifying five categories of inappropriate contents, including, toxic, obscene, threat, insult and identity hate. Then, we measured children's exposure to inappropriate comments on YouTube using the best-performing models.

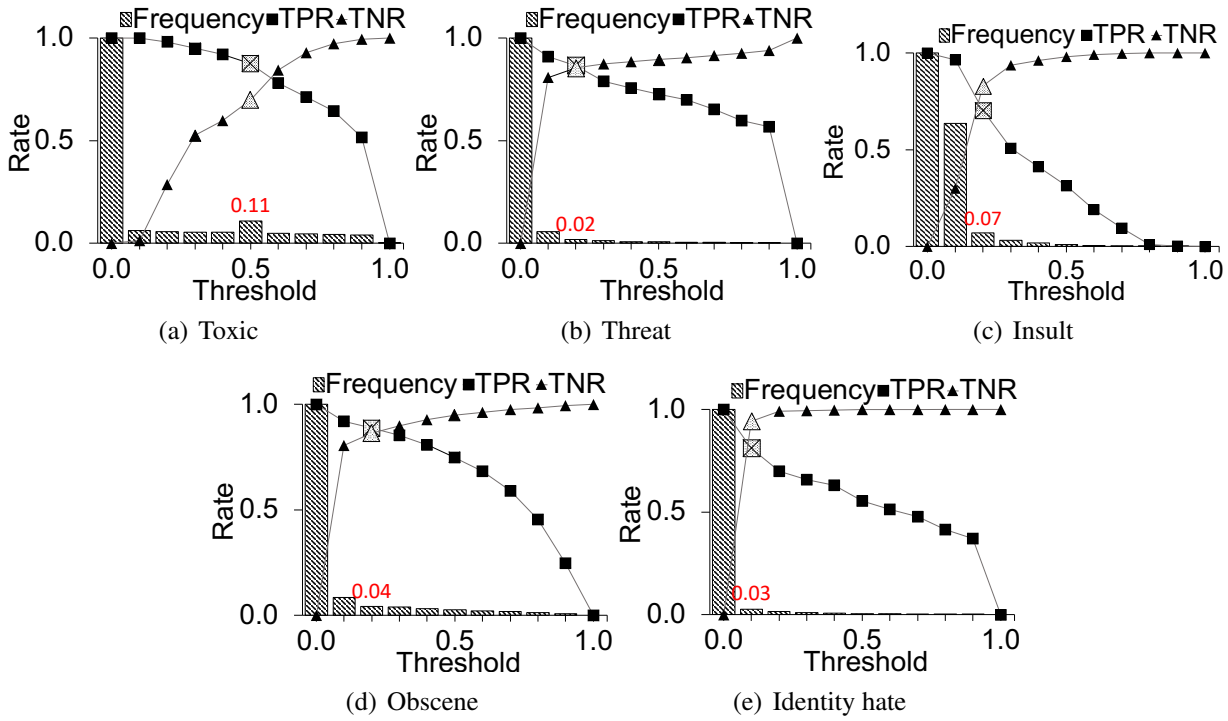


Figure 3.9: The evaluation of the ensemble model across categories in terms of TPR and TNR. The x-axis represents the chosen threshold, and y-axis shows the respective TPR, TNR, and percentage of detected YouTube comments.

Ensemble Model Performance

The ensemble model performance is reported in Table 3.2 using three metrics. We reported the performance of the models trained on Wikipedia then evaluated on the annotated YouTube comments as well as the performance of these models after being fine-tuned. The results are based on the specific probability threshold providing the best trade-off between TPR and TNR as shown in Fig. 3.9. An emphasis on high TPR is considered when choosing the threshold to ensure high correctness for positively predicted output (i.e., some positive contents might not be detected but are barely mistaken when they are detected). This high performance can be seen with the F1-score, with a high of 86.6% for the toxic and a low of 52.9% for the threat. We also observed the challenge in achieving high TPR for the threat and identity-hate categories due to several reasons, including the limited number of samples for those categories (see Table 3.1) and the ambiguity caused by the

used language.

Ensemble Adoption and Measurement

Using the best TPR-TNR trade-off thresholds, we constructed an ensemble model to evaluate and measure kids exposure to inappropriate comments. We first show the measurement using the individual models, followed by the overall performance of the ensemble of multiple models for binary classification task.

(1) Toxic Comments. We measured the toxicity of YouTube comments using the toxic comments detection model. Figure 3.9(a) shows the performance of the model in terms of TPR and TNR using different thresholds, and 0.520 is selected as the threshold with the best trade-off. Applying the model on our dataset, Figure 3.9(a) shows 11% (405,290 comments) of all comments were classified as toxic.

(2) Threat Comments. Similarly, the model for detecting threat comments achieved a TNR of 86%. We set the threshold for this category to 0.220, providing the best trade-off with a TPR of 85% as shown in Figure 3.9(b). Adopting the model to detect threat comments, 2% of the comments (63,939 comments) were labeled as threat.

(3) Insult Comments. The insult comments model provides a TPR of 66% and TNR of 85%. Figure 3.9(c) shows the results using different thresholds. In our design, we selected 0.210 as a threshold for predicting insult comments. Using this model with the adopted threshold, 7% of the collected comments are detected as an insult (262,934 comments).

(4) Obscene Comments. The obscene comments model, operating with a prediction threshold of 0.270, achieves a TPR of 86% and TNR of 88%. Figure 3.9(d) shows the results of adopting different thresholds, most of which provide high scores. Applying the model on the comments, 4% were detected as obscene (159,823 comments).

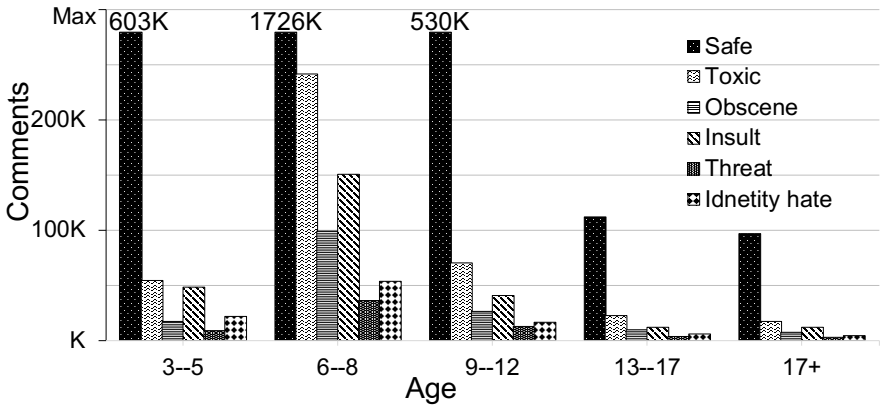


Figure 3.10: The distribution of inappropriate comments over different age groups.

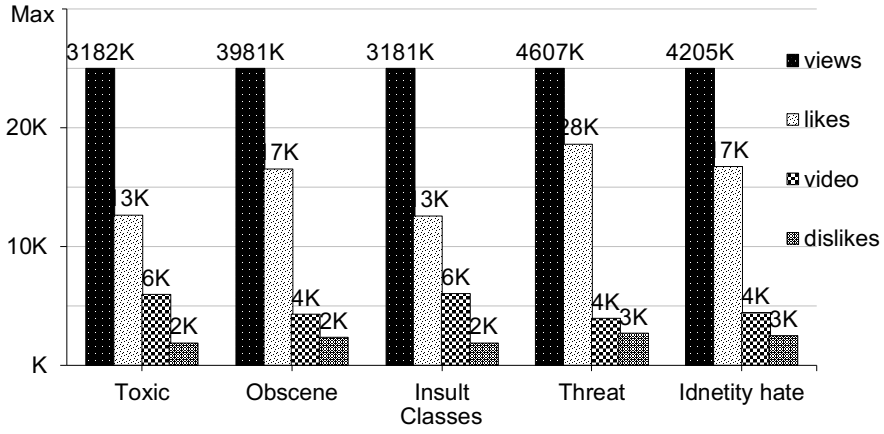


Figure 3.11: The average number of views and likes on kids' videos containing inappropriate comments.

(5) Identity Hate Comments. The model for detecting identity hate comments shows a high performance as demonstrated in Figure 3.9(e). Using a prediction threshold of 0.140, the achieved TPR and TNR are 74% and 98%, respectively. Applying the model to YouTube comments, we found that among the comments, 3% were labeled as identity hate comments, or 101,311 comments.

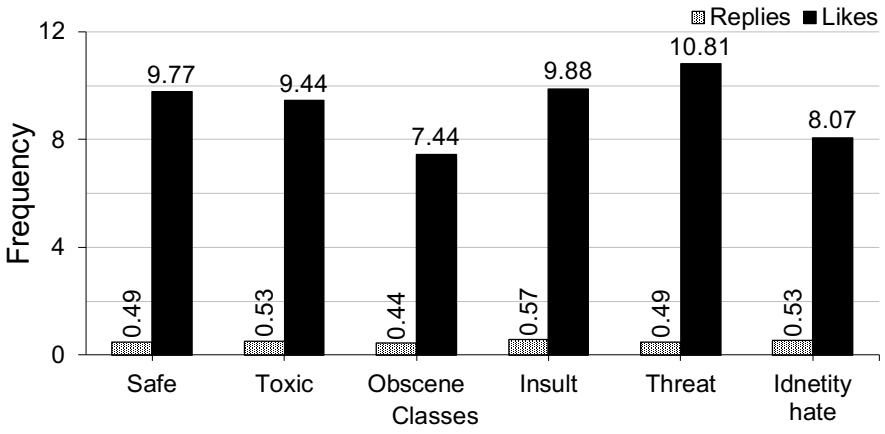


Figure 3.12: The average number of likes and replies received by the inappropriate comments.

Inappropriateness Exposure Analysis

Exposure by Age-Group. Applying the ensemble models shows the exposure magnitude of kids to inappropriate content on YouTube comments. Investigating the exposure by different age groups, Fig. 3.10 shows the distributions of the inappropriate comments from each age-inappropriate category over different age groups. For simplicity, we studied the contents of comments posted on YouTube videos targeting different age groups instead of the age in years (distributions of collected comments on videos for a specific age is shown in Fig. 3.3). Applying the ensemble models on the collected comments, we observed that toxic comments are highly common in children's videos and exceed 200,000 comments on videos only targeting the age group of six to eight years old. Insulting comments can be clearly noticed in videos targeting young children; e.g., age group 3-5 has 48,306 comments, which corresponds to 6.83% out of the total comments collected on videos of this age group (707,161 comments). Comments with some sort of toxicity are also present in the collected dataset with 81,303 toxic, 17,384 obscene, 48,306 insult, 9,065 threat, and 21,329 identity hate which were detected in comments posted on videos for the age group of 3-5. These records increase to 241,352 and 36,150 for toxic and insult, respectively, on videos for the age group of 6-8. These patterns of appearance for toxic comments are observed for videos targeting all age groups. The number of comments that contain obscene, threat, and identity hate are notice-

ably high for all age groups (e.g., they reach 99,165, 36,150 and 53,517, respectively, for the age group 6-8). We note that the reported numbers of detected categories of inappropriate comments in Figure 3.10 do not reflect their percentage with respect to the total number of comments for a certain age group. We observed that children in the age group of 3-5, which are the youngest audience, are the second most exposed to inappropriate comments, with 7.71%, 2.46%, 6.83%, 1.28%, 3.02% for toxic, obscene, insult, threat and identity hate categories (out of the total), respectively. This age group is only second to the 13-17 age group, which has 15.54%, 6.84%, 7.96%, 2.24%, 4.20%, for the same types.

Exposure and User Interaction

Acquiring YouTube kids videos, where comments were collected and investigated, is done using the top-50 search results from the YouTube Search APIs with measures of relevance and popularity (i.e., it is safe to state that the considered videos are popular). We show statistics of users' interactions with videos that contain different inappropriate content (for the five investigated categories) in Fig. 3.11. Considering the number of videos with age-inappropriate comments, we observed that the highest number of videos (6,037 videos) are reported for those with insulting comments, which has the second most number of comments among other categories (262,934). The videos with threatening comments have an average of 6.4 million views and 18,640 likes per video. More interestingly, videos with threatening comments tend to get higher interaction in terms of the number of likes (18,640) than videos with either obscene or identity hate comments (an average of 17,000 comments). Another observation is that the number of dislikes for the videos is positively proportional to the number of threat and identity hate comments.

We explored user interaction with inappropriate comments in terms of the number of likes and replies. Figure 3.12 shows the average number of likes and replies for comments that belong to the five inappropriate categories. The more likes and replies a comment gets will increase the likelihood of that comment being shown in the top comments. We have noticed that threatening

Table 3.3: Distribution of the inappropriate comments over different YouTube Channels as well as the average of the inappropriate comments to the overall comments posted on each category.

Channel Name	# Video	# Comments	Safe	Toxic	Obscene	Insult	Threat	Identity hate	Unsafe / video	Unsafe / comment
Warner Bros. Pictures	3	140594	113569	20994	10980	8002	1761	2954	9008	19.2
Cartoon Hangover	52	118352	101385	11013	4191	6625	1926	1664	326	14.3
Talking Tom and Friends	44	99293	88935	5374	460	4491	1077	2272	235	10.4
Cartoon Network	81	89620	80142	4003	137	3320	1956	1763	117	10.6
moviemaniacsDE	3	21052	11948	7012	3795	3631	595	1397	3035	43.2
Flashback FM	16	40833	29301	8788	4202	4376	940	1170	721	28.2
Mickey Mouse	46	56423	50159	2230	106	3411	561	1183	136	11.1
Nickelodeon	38	46097	42730	1222	66	1629	489	595	89	7.3
DEATH BATTLE!	3	45652	39448	3608	810	2242	1094	634	2068	13.6
Official Pink Panther	60	41730	36950	1388	175	1738	353	2197	80	11.5

and insulting comments have the highest average of likes and replies, e.g., around 11 likes and 0.5 replies per comment for the threat category. As opposed to the other age-inappropriate categories, identity hate, and insulting comments have a high number of average replies, with an average reply of 0.53 per comment. Even though the users interaction with comments from other categories is less than threatening and insulting comments, the interaction can be seen for all categories in Fig. 3.12.

Exposure by YouTube Channel

Investigating the top-10 most comment-contributing YouTube channels to our collected comments, Table 3.3 shows the distribution of age-inappropriate comments across different channels with respect to the five investigated categories. The table highlights the number of videos of which we collected the comments as well as the number of collected comments enabling the estimation of the percentages of inappropriate comments. The highest number of detected inappropriate comments is reported for *moviemaniacsDE* channel with 43.2% of the total comments classified as inappropriate. Furthermore, there is an alarming number of unsafe comments posted on the *Warner Bros. Pictures* channel videos, where the average number of inappropriate comments is 9,008 comments per video. In contrast, *Official Pink Panther* has the lowest average number of unsafe comments per video, with only 80 comments per video. We also observed a high number of

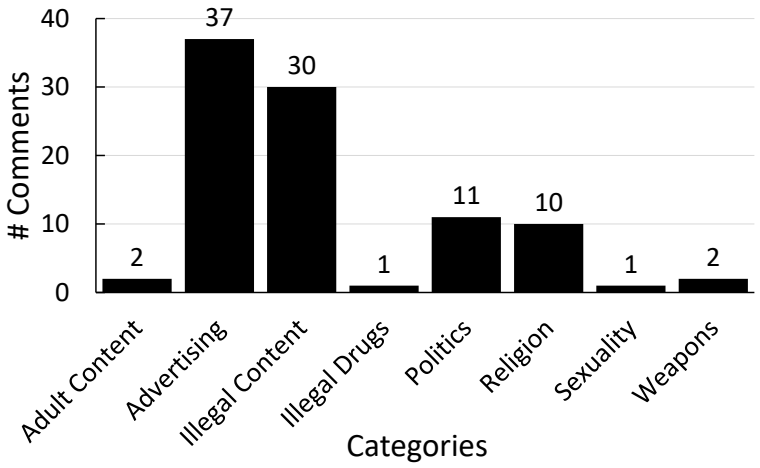


Figure 3.13: The distribution of inappropriate URLs over different IAB Categories.

detected inappropriate comments from the Nickelodeon channel, with 7.3% of the total comments in this channel classified as inappropriate. This percentage was the lowest among other channels, although still an alarming score of exposure to inappropriate comments for impressionable children.

URL Content Analysis

This study investigates how appropriate the URLs embedded in the comments on YouTube kids' videos. As previously mentioned, the audience may intentionally or accidentally access the content of the URLs, highlighting the importance of understanding the content and its effect on the children. While it is not possible to know how many users accessed the URLs, we defined two metrics to estimate the prevalence and use of the URL by the audience, including 1) video popularity, represented by the number of views, likes, and comments on the video including the URL, and 2) comment popularity, defined as the likes and replies on the comment containing the URL. While the latter does not necessarily capture the context of the engagement, it captures the level of engagement as a magnitude, which is essential in capturing users' exposure.

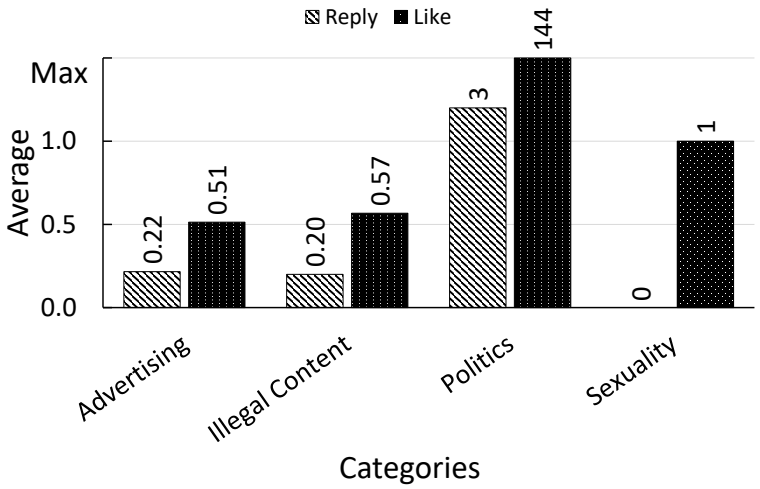


Figure 3.14: Users' interactions with inappropriate URLs from different Categories.

Kids Exposure to Inappropriate Topics

Within the 107 IAB extracted topics, eight topics are highly inappropriate for children, including "*Adult Content*", "*Advertising*", "*Illegal Content*", "*Illegal Drugs*", "*Politics*", "*Religion*", "*Sexuality*", and "*Weapons*". Note that other topics may not be appropriate for children, however, we only considered the most obvious topics that are directly inappropriate for children to be exposed to.

Figure 3.5 shows the number of URLs associated with inappropriate topics. In total, 94 URLs were classified as inappropriate, with an increasing trend in such URLs over the years, with three folds increase between 2017 and 2018. This indicates the risk of children's exposure to worrisome content that is not appropriate for their age. Figure 3.13 shows the distribution of the URLs among different inappropriate topics. While topics such as "*Advertising*" and "*Illegal Content*" are popular within the URLs, with 71.27% of the total URLs associated with these two categories.

In addition, it is important to understand the users' interaction with the URLs' comments. Figure 3.14 shows the number of likes and replies associated with four different inappropriate categories. Note that the other categories were excluded as the users did not interact with their com-

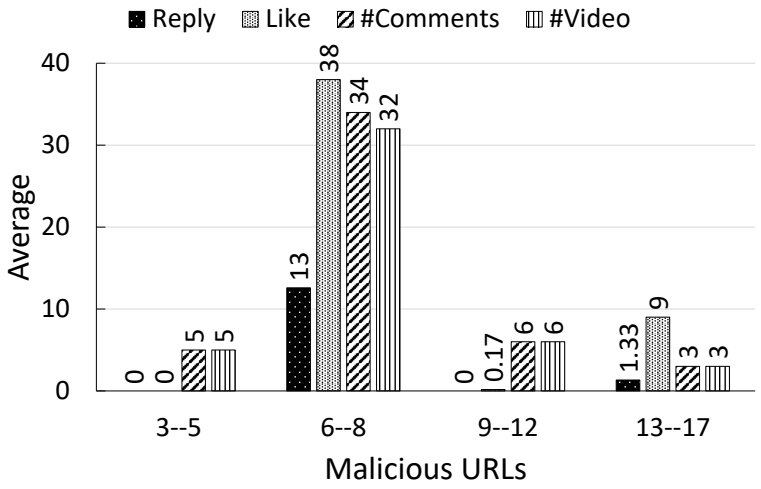


Figure 3.15: Users' interactions with Malicious URLs for all age groups.

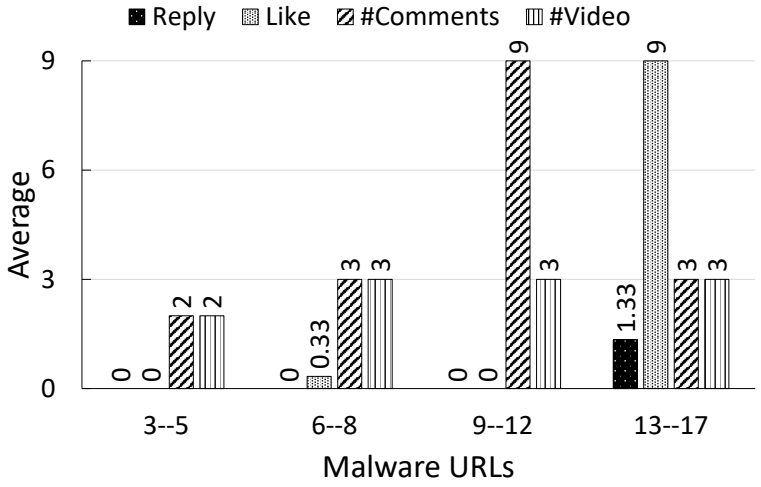


Figure 3.16: Users' interactions with Malware URLs for all age groups.

ments. As shown, comments with political URLs have on average three replies, and 144 likes, which is abnormal given the video/channel targeted audience. In general, the inappropriate URLs within the YouTube kid's comments are on the rise, leading to a potential risk of kids' exposure to their content.

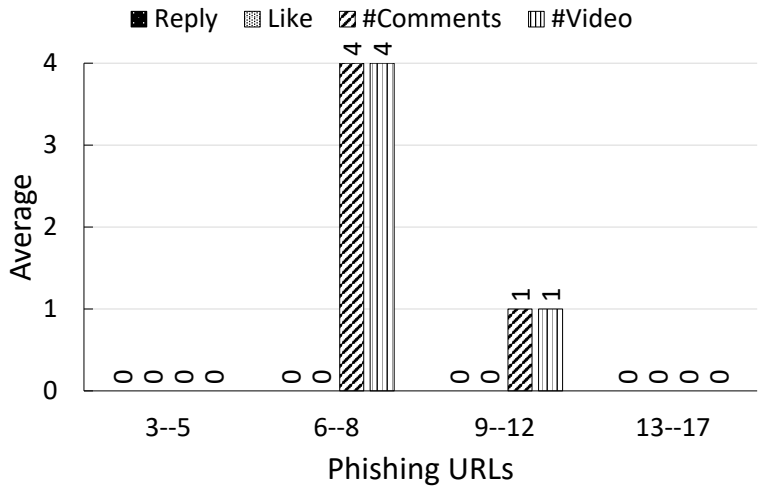


Figure 3.17: Users' interactions with Phishing URLs for all age groups.

Kids Exposure to Malicious URLs

Measuring kids' exposure to malicious URLs by different age groups, Figure 3.6 shows over years number of malicious URLs embedded in the comments. Similar to the inappropriate topics, the number of comments with malicious URLs is increasing over the years. Note that our collected dataset only includes the first three months of 2019. Figure 3.15 highlights the interaction of each age group with malicious URLs, where kids from the age of 6 to 8 have the highest interaction with malicious URLs, represented as the average number of replies, likes, comments, and videos. Furthermore, we studied the kids' interaction with malware URLs, as shown in Figure 3.16. Here, the age groups 9-12 and 13-17 show the highest interaction with malware URLs, represented with the likes and comments on the mentioned comments. Similarly, Figure 3.17 shows kids' interaction with phishing URLs. Note that only two age groups (*i.e.* 6-8 and 9-12) include phishing URLs, however, the users did not interact with their comments.

In more detail, Table 3.4 shows that videos with malicious URLs embedded in their comments have high users' interaction and engagement, which can be seen in the average number of views, comments, likes, and dislikes. There are a total of 41 videos with malicious URLs embedded in their comments, with an average of more than 46 million viewers. Based on that analysis, we can

Table 3.4: Distribution of the detected malicious URLs over different age groups as well as the average number of viewers, likes, dislikes and replies for each age group.

	Videos with Malicious URLs					Comments with Malicious URLs		
Age group	#Videos	Avg_comments	Avg_viwers	Avg_likes	Avg_dislikes	#Comments	Avg_likes	Avg_replies
3-5	7	11,218	232,743,621	330,894	125,944	7	0	0
6-8	39	20,714	21,138,213	211,227	14,122	41	142	10
9–12	10	24,395	42,455,109	206,904	18,647	16	0.06	0
13-17	4	2,203	1,380,479	14,563	806	4	9	1.33
Total	60	18,958	48,043,712	211,297	27,026	68	86	6.44
Overall[1]	9,996	650	2,106,472	8,119	1,230	3,712,911	9.87	0.50

Table 3.5: Distribution of videos and comments containing different malicious URLs as well as the average number of viewers, likes, dislikes and replies for each malicious type.

	Videos with Malicious URLs					Comments with Malicious URLs		
URLs Type	#Videos	Avg_comments	Avg_viwers	Avg_likes	Avg_dislikes	#Comments	Avg_likes	Avg_replies
malicious site	47	10,017	46,061,532	136,621	24,234	49	120	8.94
malware site	8	26,288	51,075,237	284,887	28,923	14	0.07	0
phishing site	5	91,286	61,825,765	795,507	50,234	5	0	0
Total	60	18,958	48,043,712	211,297	27,026	68	86	6.44
Overall[1]	9,996	650	2,106,472	8,119	1,230	3,712,911	9.87	0.50

safely consider these videos as popular, which would attract more users. Moreover, the videos with malicious URLs targeting kids from the age of 3 to 5 have the highest average number of viewers, with more than 200 million views, followed by the age group 9-12, with around 42.4 million views.

Table 3.5 lists the three types of malicious websites with the number of videos and comments containing each of the malicious URLs, as well as general statistics that show users' interactions with each of them. We can see that videos with malware sites URLs have an average number of viewers of more than 51 million views, which makes the possibility for a large number of people getting affected by malware much higher. The results also show that there are more than 61 million viewers of the videos with phishing URLs embedded in their comments, which is also an alarming finding in itself, since a higher number of viewers increases the likelihood of clicking on these links.

(a) Safe (b) Toxic (obscene, insult) (c) Threat (d) Identity hate

Figure 3.18: The most frequent words in YouTube comments per category. Since *toxic, obscene, and insult* share similar frequent words, we present them in one cloud.

Discussion

YouTube Platform for Children. Social media has become well-established and a part of most people's daily routine [26]. Many studies have shown that children under the age of 18 spend a substantial amount of time on social media, especially on YouTube. A survey, conducted by the Pew Research Center in 2018, shows that 81% of parents in the United States with children younger than 11 years of age allow their children to watch YouTube videos, and 34% of parents stated that their children watch YouTube videos regularly [68]. The collection of our dataset confirms the rapid growth of popularity for YouTube videos targeting children. More importantly, the results show that posted comments on children's videos contain contents that are inappropriate, this might affect their safety, privacy, intellect, emotion, or/and behavior. We note that YouTube established the *YouTube Kids* mobile app (in February 2015) and website (in August 2019), a safe platform for kids where the comment feature is disabled. However, a large percentage of children; i.e., 80% according to a study by [13], still use YouTube's original website and/or mobile app. Therefore, and based on our study, children who use YouTube unsupervised might encounter inappropriate content in the comments section, highlighting the risks of media platforms, and calling for measures to ensure their safety online.

Awareness of Inappropriate Comments. This study sheds light on the exposure of adolescents to inappropriate comments on YouTube, and shows that visual and audio are not the only media

that should be supervised but also the written contents. Fig. 3.18 shows some of the frequently inappropriate words detected to be one of the five age-inappropriate categories investigated in our study from comments posted on children's videos. This study shows that among inappropriate comments, there exists a large number of comments that have toxic, threatening, insulting, or/and identity hate contents which possibly can influence the psychological well-being of children.

Summary

In this section, we studied the exposure of kids to inappropriate and malicious content posted on YouTube videos. We studied the exposure to malicious URLs as well as the five age-inappropriate categories, namely, toxic, obscene, insult, threat, and identity hate. Using an ensemble of specialized models trained on labeled data, we measured the exposure of each category by different age groups to find out that the age group of 13-17 is the most exposed group to the inappropriate comments followed by the 6-8 age group. The results show that toxic comments are common on children's videos with 10.95% of the total comments having toxic language, followed by insults (7%), obscene (4.32%), identity hate (2.74%) and threat (1.73%) comments. We also measured users' interactions (views, likes, and dislikes) with videos having age-inappropriate comments as well as the comments themselves. We found that videos and comments with toxic or threatening comments tend to have higher interaction. Videos with threat comments have a high degree of popularity with an average of 4.6 million views and 28,000 likes per video. Similar popularity is observed for comments promoting identity hate with an average of 4.2 million views and 17,000 likes per video. Besides that, we have also studied the presence of malicious URLs embedded in comments, and the potential kids' being a victim to some sort of malicious activities. This research shows that children are exposed to inappropriate comments, and call for increased awareness of such exposure and take measures to ensure children's safety from this exposure while on YouTube.

CHAPTER 4: INVESTIGATING ONLINE TOXICITY IN USERS INTERACTIONS WITH THE MAINSTREAM MEDIA

People around the globe use social media as an essential part of their daily lives, not only for socializing with each other, but also as a major source of news. As such, and 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 users logged in monthly [88]). This huge growth has attracted news publishers to deliver their content through video-sharing platforms for fast delivery of content to viewers, and to enable the social component of interaction with their viewers, which is enabled by the comment section of videos. A recent study by Smith *et al.* [68] shows that the number of users getting their news from YouTube, for example, has nearly doubled between 2013 (20%) and 2018 (38%), and 53% of YouTube users consider YouTube as an important source for understanding current events from around the world.

A major feature of video-sharing platforms such as YouTube used for delivering news stories, as compared to the traditional medium, 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 detecting inappropriate comments [23, 32]. Despite these efforts focused on detecting inappropriate comments, the associations between various types of toxicity and topics covered in news videos from mainstream media remain an unexplored challenge. This research explores such associations through establishing automated methods for detecting different types of toxicity and topic discovery.

This work has been published at Accepted in the 5th International Workshop on Mining Actionable Insights from Social Networks (MAISoN 2020); held in conjunction with ICWSM 2020.

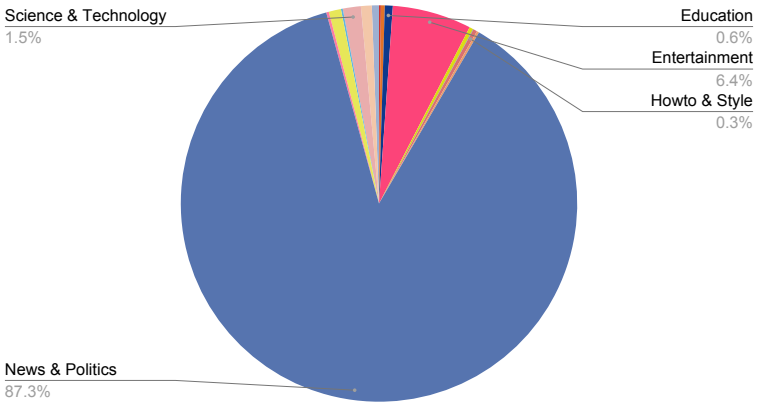


Figure 4.1: The distribution of the news videos over different categories provided by YouTube.

The goal of this work is to explore new automated methods to detect various types of inappropriate comments, including hate, obscenity, and threat comments that are posted on the mainstream news channels on YouTube. To achieve this goal, we proposed the design and implementation of an ensemble deep learning-based approach leveraging several powerful capabilities of state-of-the-art techniques on data representation and modeling to detect five types of toxic comments. We trained our ensemble model based on our manually annotated ground truth comments totaling approximately 6,000 comments. Our ground-truth augments a dataset made available dataset by Conversation AI team [14], which includes comments of the five investigated toxic categories (toxicity, obscenity, insult, threat, and identity hate). Our approach achieves remarkable results in detecting various inappropriate comments, posting, for example, an F1-score of 86.6% on detecting toxic comments. For detection, we evaluated our ensemble-based approach on a large-scale dataset of comments that includes more than seven million comments posted on more than 14 thousand videos from 30 worldwide news channels on YouTube from 2007 to 2019. Among various significant findings, the detector used in this study detected a large number of inappropriate comments posted on news videos, including 1,648,345 toxic comments that constitute 22.4% out of the total comments on the studied videos.

Effective and accurate detection of toxic comments enables enhanced online experience by moderating comments that negatively impact social interactions and public discourse. Furthermore,

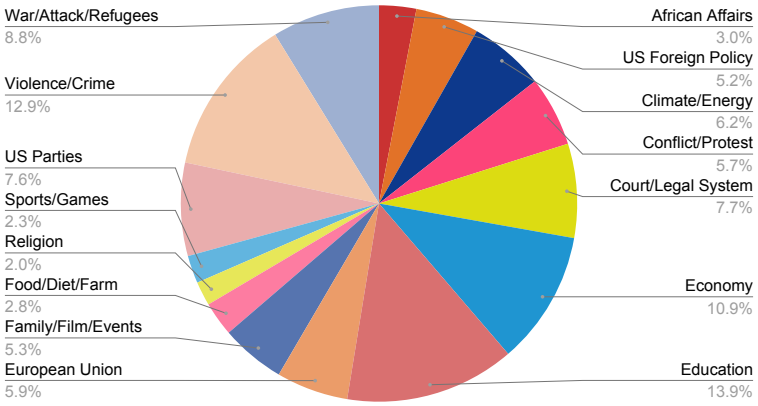


Figure 4.2: The distribution of the news videos over different generated topics using LDA.

exploring the relationship between specific types of toxicity and topics provides insights on different sources of bias and the intellectual relationship to the audience. While the root cause of the inappropriate comments posted on the YouTube videos can be attributed to multiple reasons, such as being *topic-driven* (influenced by the topic covered on the video from which the comment is collected) or *response-driven* (based on socially interacting with other comments from different users), this work provides an in-depth analysis between the relationship of such comments and the covered topics 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 covered topics. Relying on the YouTube categorization feature does not accurately capture the topics of the video. For instance, Figure 4.1 shows that 87.3% of videos published by news channels are categorized 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 (as shown in Figure 4.2) and observe the posted comments in relation to these 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 contributes to investigating the online toxicity observed in the comments posted on mainstream channels videos. Moreover, we aim to establish associations of different

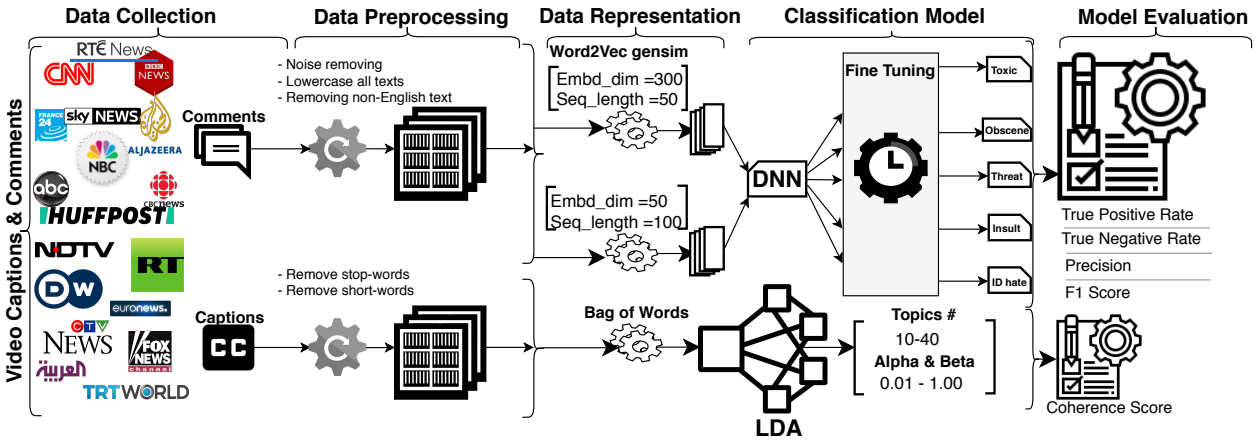


Figure 4.3: The system design pipeline. The design consists of five components, including data collection, preprocessing, and representation, followed by the ensemble classification model and topic modeling, and model evaluation and results reporting.

types of toxicity with topics covered on the news. 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 thousand news videos. We manually annotated approximately six thousand comments to five 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 popularity and users' interactions with such comments.

Data Collection and Measurements

In this section, we describe the methods we used for data collection and measurement, data preprocessing, data representation, toxicity detection, classification, and topic modeling. Figure 4.3 shows the pipeline of the system starting from the data collecting to the final step which is the evaluation. 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 comments posted 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 with the annotated ground truth dataset from the Conversation AI team [14] for the purpose of training our models.

YouTube News Channels. We collected comments on YouTube videos published by the most viewed mainstream media based on Ranker [57], a website that relies on millions of users to rank media contents such as shows and films. We extended our list of mainstream media from a Wikipedia list of the most viewed news channels [83]. In collecting channels, we were only concerned with channels that use English as the primary language in delivering their news. Both Ranker and Wikipedia lists have well-known channels that use non-English speakers to broadcast their news; those channels were not included in our final list, which included 30 channels.

Collection Approach. Using YouTube APIs, we gathered the most popular 500 videos posted on each of the selected channels (the 30 most ranked channels that constitute our list). We observed that some channels have less than 500 videos, and therefore the final list of targeted videos contains 14,506 videos. In the collection process, we utilized four different YouTube APIs [87], each is used for obtaining a different set of features. We retrieved videos IDs using *YouTube search API*, which takes channel name as an argument along with a filtering option that is set to the popularity of the video. After passing the search argument, the API returns a list of videos IDs, which we used to extract the comments. To extract comments and statistics (*e.g.* number of views, likes, dislikes)

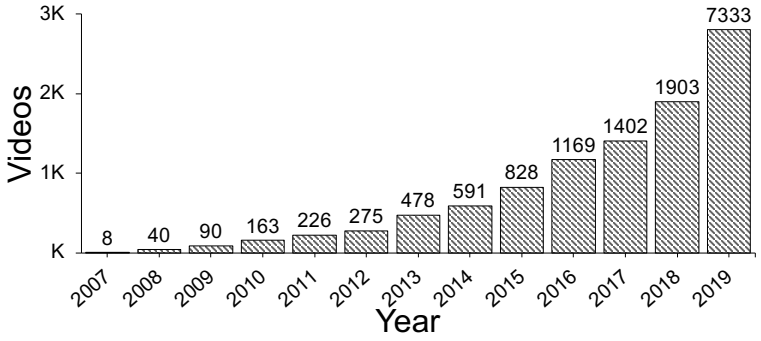


Figure 4.4: The distribution of news videos over the past years.

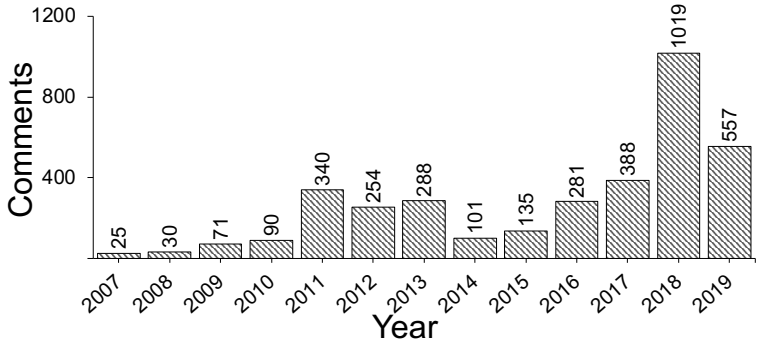


Figure 4.5: The average number of comments per video.

of each video ID, we used two different APIs, namely, *YouTube comments* and *YouTube videos*. To obtain the caption of videos, we used *YouTube transcript API*. By doing so, we collected more than 7.3 million comments from 14,506 videos, and 10,883 video captions.

Data Statistics and Measurements. The popularity of the selected channels can be seen in the average number of 1.8 million subscribers for each channel. Also, the videos uploaded by these channels have a high average number of views and comments, with 383,402 views and 509 comments per video. We collected a total of 7.3 million comments posted by 2,992,273 unique users. Figure 4.4 shows the distribution of the observed videos over the past 13 years (2007 to 2019). As shown, most of these videos were published in 2019, as the trend shows an increase in videos popularity in recent years.

Moreover, we studied people's interactions with the videos through comments, and Figure 4.5

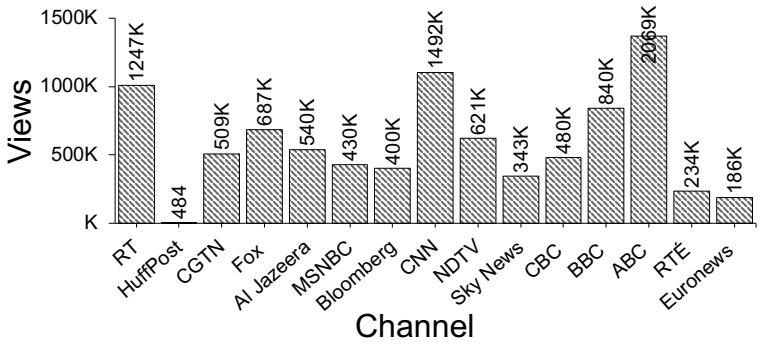


Figure 4.6: The average number of views per news video for the top-15 mainstream channels.

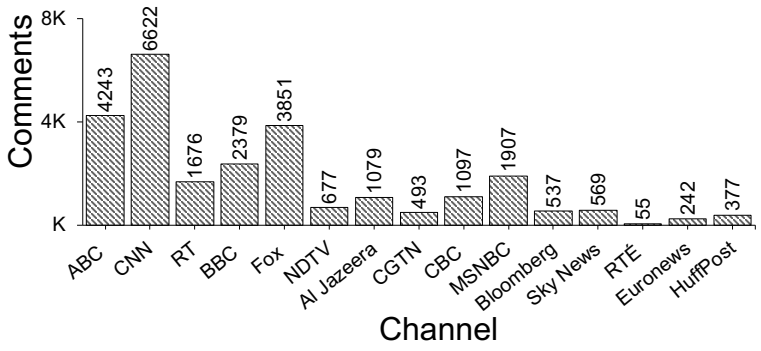


Figure 4.7: The average number of comments per news video for the top-15 mainstream channels.

shows the average number of comments on each video across different years. We found that the videos published in 2018 maintain the highest average number of comments with 1,019 comments per video. Another factor showing the popularity of the channels used in our study is the average number of views, as shown in Figure 4.6 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 4.7. 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. Studying user's behavior in the comment section, including toxic behavior, we utilized two datasets to train a machine learning-based ensemble classifier for toxic comment detection and classification: (1) Wikipedia comments created by Conversation AI team [14] and (2) our own manually annotated YouTube comments.

❶ **Wikipedia Ground Truth Dataset.** This dataset has around 160,000 comments from Wikipedia Talk pages, manually annotated by the Conversation AI team [14]. In this dataset, around 143,000 comments are labeled as safe, while the remaining samples are assigned labels of five types of toxicity (*i.e.* 15,294 toxic, 8,449 obscene, 7,877 insult, 478 threat, and 1,405 identity hate comments). The collected data is summarized in Table 3.1.

❷ **YouTube Ground Truth Dataset.** In addition to the Wikipedia ground truth dataset, we created our own dataset by manually annotating 5,958 YouTube comments. The annotated comments are distributed as follows: 1,832 safe comments, 4,126 toxic comments, 2,367 obscene comments, 1,650 insult comments, 550 threat comments, and 788 identity hate comments. Each comment is either toxic or safe. The toxic comments may then be mapped to five different toxic categories (*i.e.* toxic, obscene, insult, threat, and identity hate). When manually labeling these comments, we considered some explicit rules that we briefly describe in the following. We labeled comments as obscene when they are morally offensive and/or when they have socially profane words. Furthermore, when such profanity is used with others, we labeled the comments as an insult. If offensive language was used against a group of people, based on their race, color, or ethnicity, the comments are labeled as identity hate. We note that, labeling comments as identity hate is a challenging, and highly subjective task [50] [61]. Some comments did not have any profanity, but they imply threats to other users or the video publisher. Therefore, we labeled such comments as a threat. Toxic comments that are socially unacceptable, and do not imply any of the aforementioned categories, were labeled as *toxic* only. We avoided using multiple annotators across different folds of the manually-labeled dataset, and rather pursued this slow labeling method, to avoid inconsistency and subjectivity in interpretation.

Safe Ground Truth Dataset. For the safe comments, we combined the safe comments of both datasets used in this study, *i.e.* our manually annotated YouTube comments and the Conversation AI team dataset of comments.

Data Preprocessing

We adopted several preprocessing steps before further data representation and modeling to clean and ensure proper handling of data. We describe the used preprocessing steps for both captions and comments as follows. ❶ We initially removed all *non-English contents* across all datasets, and limit our analysis to English-written comments. ❷ We eliminated irrelevant characters and tokens, *e.g.* punctuation, and other characters that represent or encode emojis. ❸ We removed the stop-words for all datasets. Furthermore, we conducted additional steps to preprocess the captions including the following. ❹ We removed frequent words that appear in more than 50% of the news scripts as they are shown to be for describing general contents. We also eliminated the rare words that appear less than five times in all the corpus. ❹ We applied both lemmatization and stemming process, the lemmatization done using WordNet lemmatizer [45] and the stemming process, using the Snowball stemmer [69].

Comments Data Representation. To perform a machine learning task on textual data, the text has to be transformed into a numerical representation. This process allows the machine learning models to learn and capture various patterns of the text.For the numerical representations processing, the raw text is inputted and transformed into numerical vectors that vary in shape and size based on the utilized representation method. For data representation, we explored the effects of using different data representation methods, namely, *Word2Vec* [44] and *Glove* [54]. These two methods are among the most commonly used approaches for textual data representation. Both *Word2Vec* and *Glove* are representation methods that generate word embeddings, which are numerical vectors with predefined dimensions. These word embeddings are concatenated to construct the full-text representation as a sequence of vectors with the same size as the included words. This study uti-

lized the pre-trained word representation models available with the Natural Language Processing (NLP) tools. Specifically, we use the pretrained *Word2Vec* from Gensim [59] and *Glove* from Stanford NLP Group [54].

Word2Vec. *Word2Vec* is a model trained in order to map words to numerical vectors. In *Word2Vec*, words occurring in a similar context will be mapped into similar vectors representation. *Word2Vec* is a widely used representation method in the NLP field due to its their ability to capture different relationships among words. 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. Those predictions can also be used to describe a word association with other words in the same context. (*e.g.* what "car" is to "automobile" what "man" is to "woman"). *Word2Vec* is trained using two approaches. The first approach is a continuous bag of words (or *CBOW*), which uses the input context to predict a target word. The second approach is *skip-gram*, which uses a word to predict a target context. *Skip-gram* is widely used due to its ability in capturing and producing accurate results on large datasets. With the *skip-gram* approach, we are given a corpus of words w and their contexts c. We considered the conditional probabilities $p(c|w)$, and given a corpus $Te \times t$, the goal becomes to set the parameters θ of $p(c|w;\theta)$ in order to maximize the corpus probability using the following formula: $\arg\max_\theta \prod_{w \in Te \times t} \left[\prod_{c \in C(w)} p(c|w;\theta) \right]$,where $C(w)$ is the set of contexts of word w. Using *Word2Vec*, we transformed all comments into numerical vectors. During this task, we tried different configurations (*i.e.* vector size) of the pre-trained models, and chose the one that achieve the highest accuracy during the evaluation of our models.

Glove. *Glove* is an unsupervised learning algorithm that generates numerical representations for words in a given document. The training occurred on aggregated global word-word co-occurrence statistics from a corpus. Given a corpus with V words, the co-occurrence matrix X will be a matrix of $V \times V$ elements, where the element indexed by the i_{th} row and j_{th} column of X, namely X_{ij}, represents how many times the word i has co-occurred with the word j, represented as $\frac{P_{ik}}{P_{jk}}$, where $P_{ik} = \frac{X_{ik}}{X_i}$. Here, P_{ik} represents the probability of the word i occurring with the word k, which

can be calculated by dividing the number of times both words occurred together X_{ik} over the total number of times the word i occurred in the entire corpus; *i.e.* X_i [54].

Using word representations, *i.e. word2Vec* and *Glove*, comments are represented as a sequence of word embedding as $\text{Comment}_i \in \mathbb{R}^{n \times d}$, where n is the number of words in the i-th comment and d is the dimension of the word representation. For efficient processing and tensor-based operation, *i.e.* batch of samples in predefined tensor with the dimensions of $batch\_size \times sequence\_size \times word\_embeddings$, we adopted a predefined sequence size to represent a comment. With the sequence size, comments with larger size are truncated, while smaller ones are padded with zeros.

In the following, we briefly describe the two used pre-trained models for comments representation. ❶ **Word2Vec settings:** For our experiments, we explored different configurations for representing comments using *Word2Vec*, including the word embeddings size and the sequence size (*i.e.* the number of words included for representing a comment). The best settings for comment representations using *Word2Vec* is using word embeddings of size 300 and a sequence size of 50 words. ❷ **Glove settings:** For our experiments, we explored different configurations for representing comments using *Glove*, including the word embeddings size and the sequence size. We achieved the best results by adopting word embeddings of size 50 and sequence size of 100 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 a bag of words method, in which words are assigned a unique identifier. In the bag-of-words scheme, a video caption is represented with statistical values calculated for each word of the captions. Considering large-scale datasets, the bag-of-words representation vector can be sparse and high-dimensional. To optimize our bag-of-

words representation, we adopted several steps including: (1) removing frequent and rare words, and (2) feature selection. Common words that appear in more than 50% of the captions can be very general terms with less discriminative features than other words that are less frequent. Rare words, that appear less than five times in the entire corpus, are also eliminated since they do not have abstract meaning. To reduce the dimensionality of the bag-of-words representation, we selected the most frequent 10,000 words to be the caption data representation.

Toxicity Detection Models

We started by detecting and classifying different toxic behaviors in the comments to further examine their association with the topics covered in the news of which the comments are collected. We inspected comments for five categories of toxicity: *toxic*, *obscene*, *insult*, *threat*, and *identity hate*. We utilized a deep learning-based ensemble classifier of five specialized models for classifying the five toxic categories.

Deep Learning-based Architectures. To build the ensemble, our preliminary experiments included the examination of capabilities for three deep learning-based architectures—Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN))—in capturing the toxic behavior from comments. Our preliminary results showed that DNN-based models outperformed other architectures in terms of identifying different toxicity types. Therefore, we selected the DNN-based architecture as a baseline to construct our ensemble classifier.

DNN-based Models

DNN is a supervised learning method that can discover both linear and non-linear relationships between the input and the output. We utilized specialized DNN models, each trained and evaluated to detect one of the toxic categories, and use them collectively to build our ensemble.

Comments that are represented as a sequence of word embeddings are fed to the DNN-based models to be inspected for the different underlying toxic categories. If the comment generates a negative output in all models, it can be considered as a safe comment (*i.e.* the comment does not imply toxic content). The DNN model architecture used in this study consists of (1) the input layer of size either (50×300) for the *Word2Vec* representation or (100×50) for the *Glove* representation, (2) two fully connected hidden layers of size 128, with ReLU activation function, and (3) the output layer with one unit of sigmoid activation function. This unit signaling the probability distribution of assigning input data to the targeted category. The output layer consist of one unit that applies a sigmoid activation function $sigmoid(z) = 1/(1 + e^{-z})$. The output of the sigmoid function, $\{y \in \mathbb{R} \mid 0 \leq y \leq 1\}$, indicates the probability of assigning an input data to the targeted category.

Decision Threshold. Since the output of the model is a probability distribution, different thresholds can be explored to produce the best results. A commonly-used threshold for assigning a class to an input data is 0.5 such that $\bar{y} = 1 \; if \; y \geq 0.5$). We examined different thresholds for each model detecting one of the toxic categories to get the best setting in terms of true negative rate and true positive rate.

Data Representation Effects. our preliminary experiments also included studying the effects of different data representations on the ensemble performance. We observed that different data representations impact the detection of different toxic categories. For example, *Glove* performs well for detecting threat comments, while and *Word2Vec* performs well for detecting toxic, obscene, insult, and identity hate categories).

Dataset Handling and Splitting. Using the two ground truth datasets, *i.e.* Wikipedia ground truth comments and YouTube ground truth comments datasets, we adopted two different approaches to split the datasets for training and evaluating the models. ❶ 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. ❷ We also used 50/50 training/testing splits of the Wikipedia dataset for exploring 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. To make sure that there is no data leakage between the training and testing of models, we report the results based on the performance of individual models separately.

Model Settings and Evaluation

The training was done by minimizing the binary-cross-entropy as follows: $\text{loss}(\theta) = \frac{-1}{N}\sum_{i=1}^{N}[y_i \times \log(p_i) + (1-y_i) \times \log(1-p_i)]$, where p_i is the conditional probability $p(y_i|x_i,\theta)$ for a target y_i given an input x_i and a set of parameters θ, i is the i-th record, and N is the total number of records in the training set. We used *RMSprop* as our training optimizer, which is a stochastic optimization algorithm, with a learning rate of 10^{-3} without decaying over time. We used a mini-batch approach with a batch size of 128, and a dropout regularization with a rate of 0.2. We set the termination criterion to be the number of training iterations, which is set to 100 for all trained models.

Evaluation Metrics. To evaluate the models, we used four evaluation metrics, namely, *Precision*, *F1-score*, *True Positive Rate* (TPR), and *True Negative Rate* (TNR). The Precision shows how good a model is in predicting the positive class (P = TP/TP+FP). The F1-score is the harmonic mean of the precision and recall, and is expressed as (F1-score $=$ 2TP/(2TP+FP+FN)), where TP, FP, and FN represent True Positive, False Positive, and False Negative, respectively. The TPR represents the proportion of how many positive predictions were actually correct (TPR $=$ TP/TP+FN). The TNR is the proportion of the negative predictions out the total of negative-labeled data (TNR $=$ TN/TN+FP).

Topic Modeling using LDA

Topic modeling is a statistical model used to extract a set of topics that occur in a group of documents. Topic modeling is an unsupervised machine learning technique that processes a set of documents and detects word and phrase patterns across documents to cluster them based on their similarities. The topic modeling method defines a set of predetermined topics and assigns topics to documents based on their contents [36, 9]. In our study, we used the Latent Dirichlet Allocation (LDA)-based topic modeling approach, one of the most widely used topic modeling methods. The intuition behind the LDA is to map each document in the corpus to a set of topics, each represented as a cluster of similar and related words [10].

Fine-grained Topics Extraction

We studied the associations between a specific toxic behavior (*e.g.* obscenity, insult, threat, and identity hate) and an extracted topic from videos of mainstream channels. To do so, we use topic modeling to assign topics to videos based on their caption. This is a challenging task for two reasons: ❶ YouTube categorization is generic and lacks specification of topics covered in the video script. We observed that most videos (87.3%) published by the news channels are categorized as *News & Politics* as shown in Figure 4.1. Based on our analysis of topics that appeared in news videos, Figure 4.2 shows a variety of topics including war/attack/refugees, violence/crime, sports/games, economy, and others. ❷ We observed that most videos published by news channels cover more than one news segment.

Using the LDA-based model, we extracted topics on a fine granularity of the video caption to uncover all topics discussed in each video. To this end, we applied topic modeling on caption segments of 1,000 words generated from a fine-grained segmentation process of the caption. We chose to segment captions by 1,000 words based on our analysis of the caption size, since the majority of collected captions (more than 53%) have at least 1,000 words. This segmentation ensures the

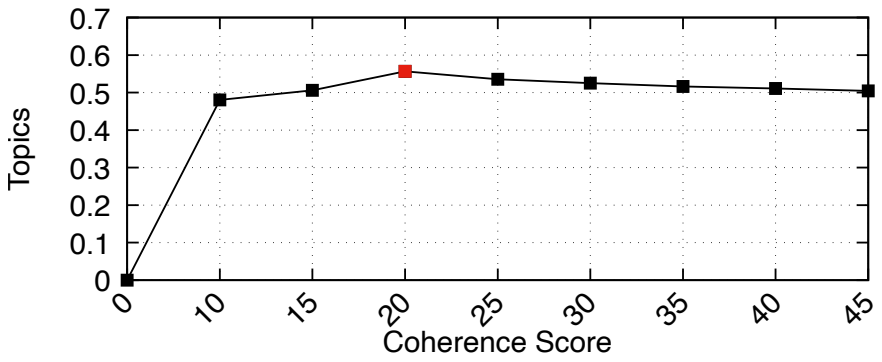


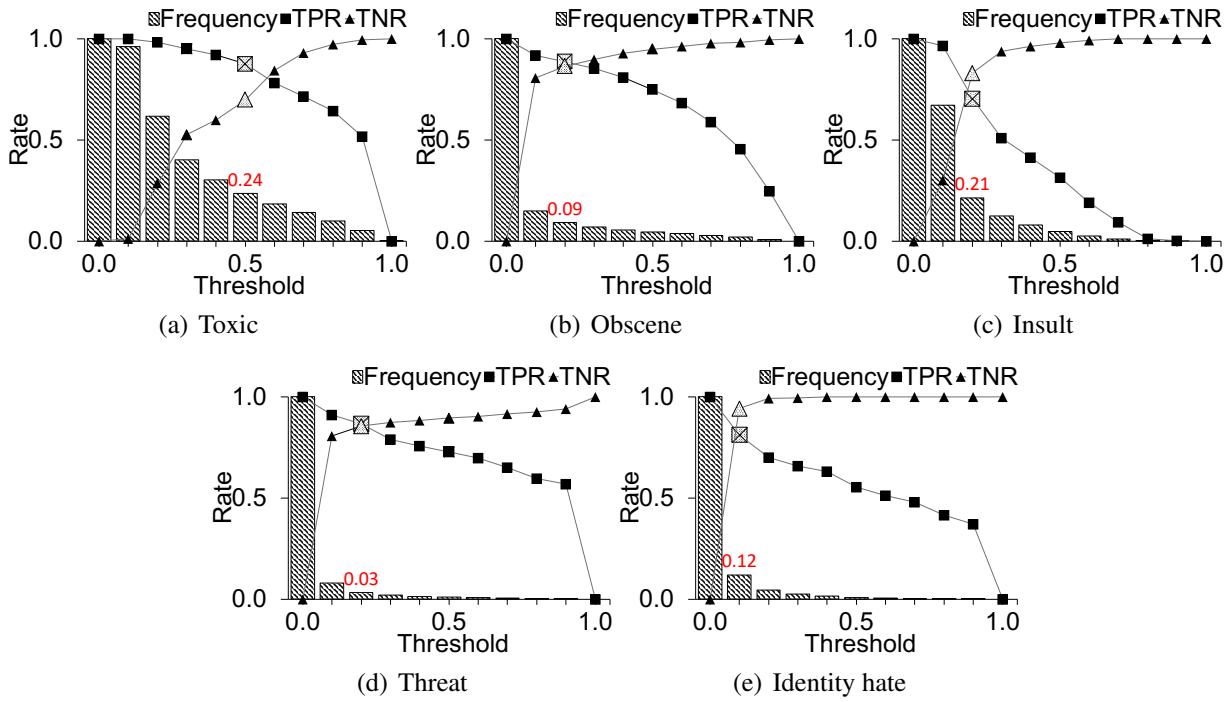
Figure 4.8: The coherence score for number of topics from 10 to 40 using alpha of 0.61 and beta of 0.31.

generation of at least one segment per caption and unveiling at least one topic. Intuitively, larger videos cover multiple topics.

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 as shown in Figure 4.8.

Using the best-performing LDA model, we manually inspected the generated topics and assigned a describing name to those topics. Since some topics share similar keywords, we assigned the same



(a) Toxic (b) Obscene (c) Insult

(d) Threat (e) Identity hate

Figure 4.9: The evaluation of the ensemble model across categories in terms of TPR and TNR. The x-axis represents the chosen threshold, and y-axis shows the respective TPR, TNR, and percentage of detected YouTube comments.

topic name for multiple extracted topics. This manual inspection of generated topics produced 15 distinct topics shown in Figure 4.2. The figure shows that the videos are fairly distributed across topics as opposed to the categorization provided by YouTube shown in Figure 4.1.

Results and Discussion

This section presents the toxicity detection in comments and provides analysis of toxicity/topics associations.

Toxicity Detection and Measurement

Using the best TPR/TNR trade-off thresholds, we built an ensemble model to detect and classify toxic comments on YouTube videos. Figure 4.9 shows the performance of models in detecting toxic

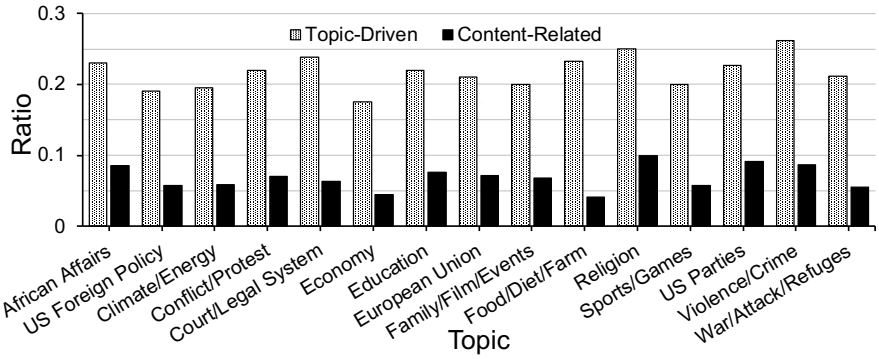


Figure 4.10: The distribution of the toxic comments over different topics generated by the LDA model.

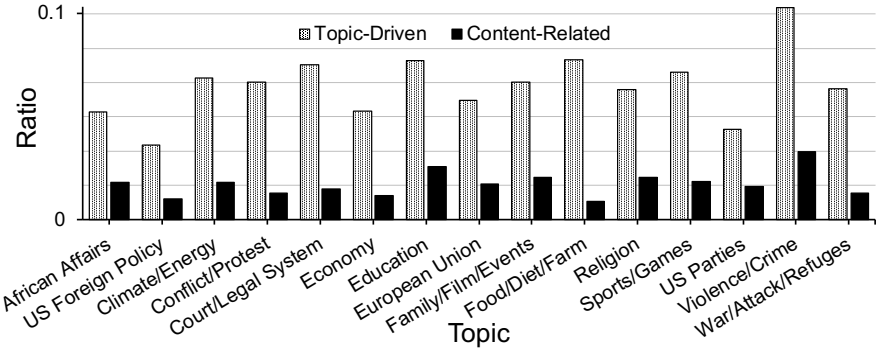


Figure 4.11: The distribution of the obscene comments over different topics generated by LDA.

comments posted on news videos from mainstream media. ❶ **Toxic Comments:** Figure 4.9(a) shows the performance of the toxic-behavior detection model in terms of TPR and TNR using different 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 with the specified threshold shows that 22.4% of the comments are classified as toxic with a total of 1,648,345 comments. ❷ **Obscene Comments:** The model operating with a decision threshold of 0.27 achieves a high TPR of 86.6% and TNR of 88.8% for detecting obscene comments. Figure 4.9(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 **Insult Comments:** The specialized model achieves a TPR of 66.7% and TNR of 85.9% when selecting 0.210 as a threshold for detecting insult comments.

Figure 4.9(c) shows the results using different thresholds. Using this model with the selected

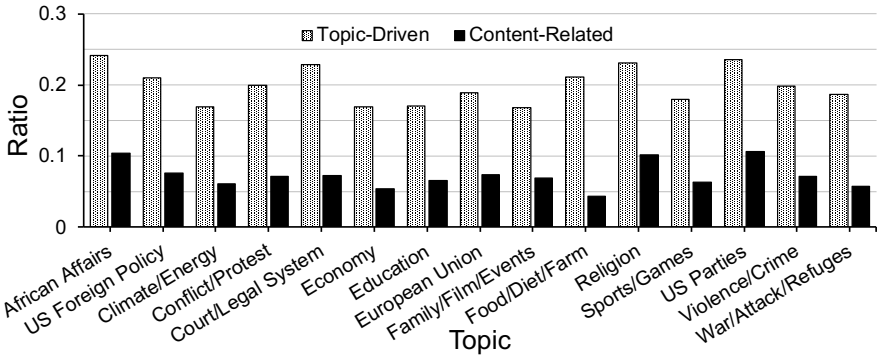


Figure 4.12: The distribution of the insult comments over different topics generated by LDA.

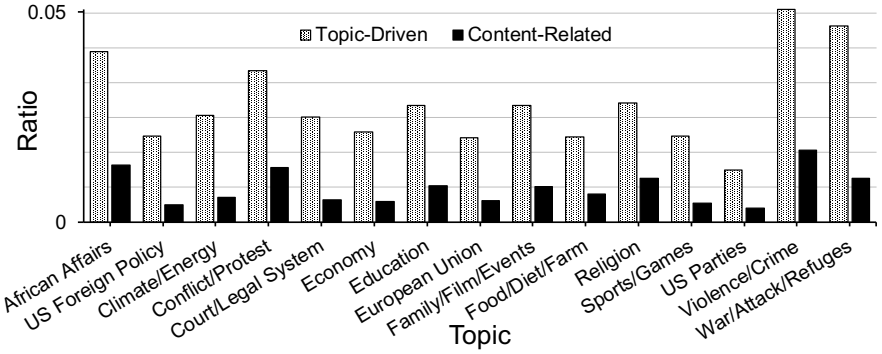


Figure 4.13: The distribution of the threat comments over different topics generated by LDA.

threshold, 19.9% of the collected comments are detected as an insult (1,465,030 comments). ❹ **Threat Comments:** The model for detecting threat comments achieved a TNR of 86.1% and a TPR of 85.4%. This result is realized with a threshold of 0.220 as shown in Figure 4.9(d). Adopting this model shows that 2.88% of the comments (211,921 comments) is classified as a threat. ❺ **Identity Hate Comments:** Figure 4.9(e) 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.

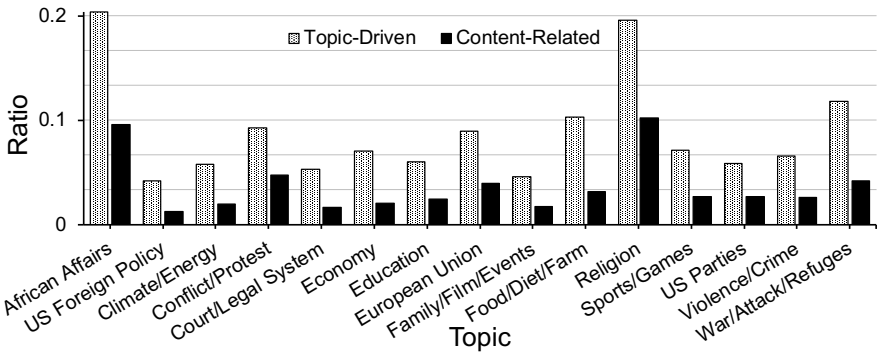


Figure 4.14: The distribution of the identity hate comments over different topics generated by the LDA model.

Toxicity and Topics Associations

The detection of several toxic behaviors and accessibility of topic categorization of videos enables us to examine toxicity/topic associations. 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 the five types of toxic comments on each topic of our LDA-based topic model. ❶ **Toxic Comments:** Figure 4.10 shows that the videos discussing topics related to religions, 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 show the highest number of obscene comments with 10% of the total comments. News about the United States foreign policy had the least number of obscene comments with only 3% of the total comment on the topic as shown in Figure 4.11. ❸ **Insult Comments:** We examined the presence of insult comments on different news topics. The news related to religion, African affairs, and the United States political parties show that about 23% of their comments are labeled as an insult. In contrast, comments posted on news related to family, education, and economy have the lowest rate of insult comments as shown in Figure 4.12. ❹ **Threat Comments:** When studying the presence of threat comments on different news topics, we observed that news related to violence/crime and war/attacks/refugees show the highest number of threat comments as shown in Figure 4.13 with

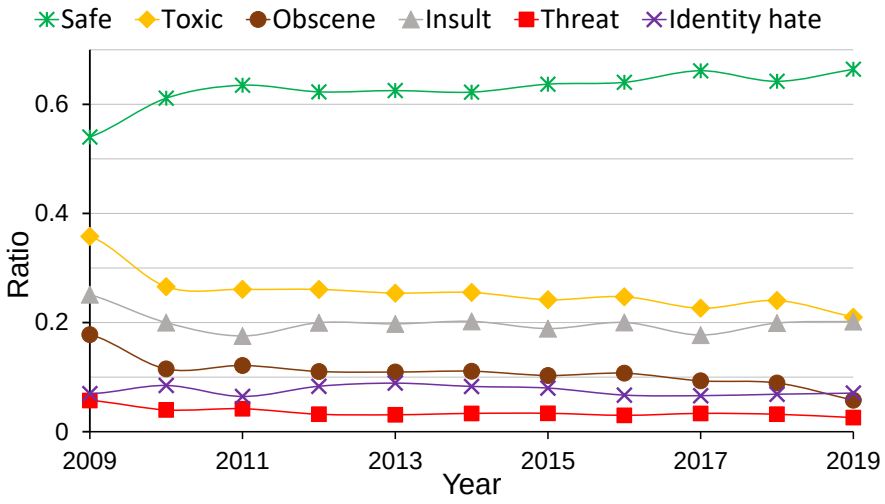


Figure 4.15: The overall ratio of toxic comments on mainstream media channels videos over the past years, maintaining roughly the same ratio over the past years.

rates of 5%, 4%, respectively. The news related to the United States political parties has the least number of threat comments with only 1% of comments. ❺ **Identity Hate Comments:** Among the 15 news topics, African affairs, and religion news have the highest ratio of identity hate comments with a rate of 20% of total 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 4.14.

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 topic's 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. We show these ratios in the figures 4.10, 4.11, 4.12, 4.13, 4.14.

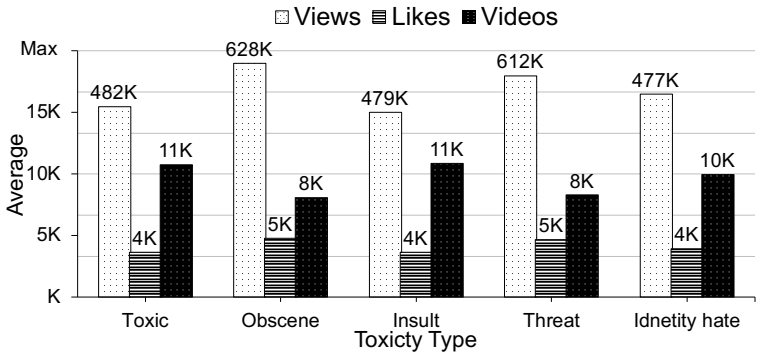


Figure 4.16: The average number of views and likes on news videos containing toxic comments. Such videos have high average number of views, likes, and dislikes.

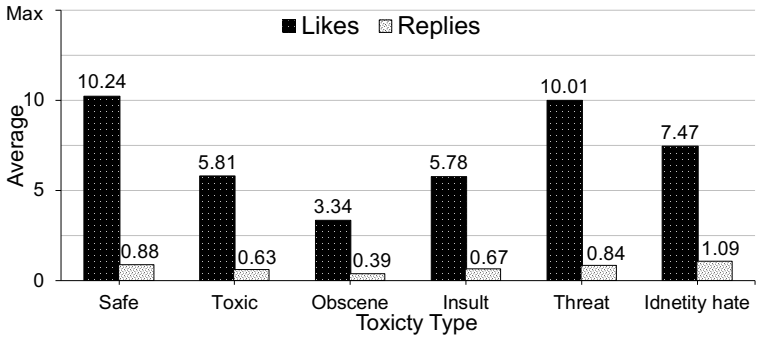


Figure 4.17: The average number of like and replies received by toxic comments. Comments associated with threat and identity hate have higher number of likes and replies.

Video and Toxicity Popularity

Despite the efforts taken by comments moderators on YouTube, we uncovered a large portion of toxic comments posted on different news topics. This seems to be almost consistent over the years as shown in Figure 4.15. This study also investigated the popularity of videos with toxic comments and the interactions of users with such comments. Figure 4.16 shows that there are more than 10,000 out of 14,506 ($\approx$69%) videos containing toxic comments. The figure shows that most type of toxic behavior is identity hate and insult.

Videos with toxic comments are popular, with a very high average number of views, where most of them have around half a million views with an average of 5,000 likes per video. Such statistics

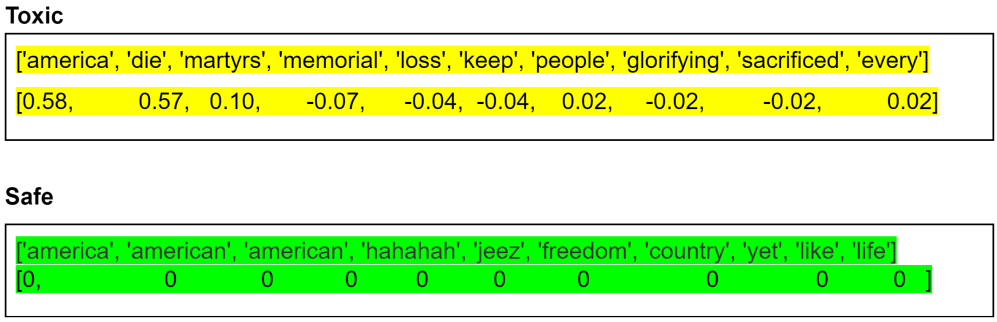


Figure 4.18: The output produced by the LIME framework for two comments with the same identity word each assigning the identity words different weight based on the given context.

indicate that more users are exposed to such toxic behavior. Moreover, we investigated the interactions with these toxic comments, these are a measure to boost the comments' popularity to become visible to a large number of users. To perform such analysis, we calculated the average number of likes and replies to comments that belong to each of the five toxic categories, as well as the safe comments as shown in Figure 4.17. Comments categorized as threat or identity hate receive high user interactions with almost one reply on average, and 10.01 and 7.74 likes, respectively.

Examining Identity Bias

In order to validate the results of the models, we performed an analysis to examine whether the utilized models in this study exhibit bias towards certain group of people. Towards that, we collected a list of identity words from different sources [60, 51, 81] and filter the toxic comments that have any of the identity word in the list. Then, we use explainable AI to pinpoint the exact word that contributed positively to the final output of the classifier. Specifically, we use Local Interpretable Model-Agnostic Explanations (LIME) which provide qualitative understanding of the model's decisions. We perform such a step to check whether an identity word contributed positively to the toxicity decision.

Using LIME on each comment, a set of words will be generated as shown in Figure 4.18 noted as *lime words* which represents the top 10 words that have an impact on the classifier decision.

Alongside the set of words, a set of weights note as *lime weights* indicate the amount of the impact for each corresponding words in the *lime words* will be generated as well. Using both the words and the assigned weights, the results shows that, there is no clear evidence that any of the identity words we investigated have by itself a positive impact on the model decision. For example, when considering the identity word *america* in a safe context the model will not assign any weight to it considering it as a toxic feature. Meanwhile, when an identity word occur in toxic context the model will assign weight to both the identity word and the toxic words. This indicates that the model is not considering some word are toxic unless they are related to identity word then it consider it an insult which is a toxic text.

Summary

In this study, we designed and evaluated an ensemble of models to detect various types of toxicity in comments posted on the mainstream media channels YouTube. We started by collecting a large dataset of YouTube comments (more than seven million YouTube comments) posted on 14,506 YouTube news videos. Exploring several deep learning architectures and experimental settings, the proposed ensemble was able to detect toxic comments with high accuracy. Despite the efforts in comment moderation taken by YouTube, our study shows the existence of a large number of toxic comments. Moreover, the average number of toxic comments is nearly consistent over the past 10 years. We also found that $\approx$69% of the collected videos contain toxic comments, with most of them were either expressing some sort of identity hate or insulting other users. Furthermore, we investigated the correlation between the content of news videos and different toxic behaviors on YouTube, namely, toxicity, obscenity, insulting, threatening, and identity hate. Using LDA on the news captions of 10,883 news videos, we extracted 15 topics, that are used to categorize the collected videos. We provided an in-depth analysis of topic/toxicity associations. Our analysis shows that religion and violence/crime-related news has the highest rate of toxic comments, while economy-related news has the lowest rate of toxic comments. This study shows the need for more

effective approaches to keep the comment section clean, especially in controversial topics that seem to draw toxic behavior.

CHAPTER 5: STUDYING USERS BEHAVIOR BEFORE AND DURING COVID-19: A MEASUREMENTS OF TRENDING TOPICS AND SENTIMENTS

COVID-19 is a highly transmissible respiratory virus and the biggest public health concern of this century, declared as a global pandemic by the World Health Organization on March 11, 2020 [84]. As of September 2021, there are more than 225 million confirmed COVID-19 cases, with 4.63 million deaths and 201 million recovered cases worldwide [85]. While older people are more likely to be at risk of serious health impact caused by the virus, young people may suffer its long-term social and economic impacts [40]. Such impacts might affect people's life and behavior, including their interaction with each other, communication, as well as their view of certain critical issues such as jobs and overall mental health [11].

The advancements in technologies, especially the major role that computer technologies play, have clearly shown their contribution to some critical medical decisions, such as pandemics and infectious diseases [8, 37, 6]. Recently, the massive data that has been gathered throughout the past years has been utilized by many researchers to enabling them to make informed decisions and disease prevention methods [52, 74]. Nowadays, data obtained from different social media platforms have become the basis for researchers and health officials for studying a variety of social and mental issues [56, 22].

Since early 2020, when COVID-19 started spreading worldwide, governments instantiate several measures in an attempt to limit the spread of the virus including, social distancing and other precaution methods were taken. This led to massive growth in using the popular social media platforms, particularly the main medium for communication and entertainment such as Twitter, Facebook

This work has been published at Accepted in the 9th International Conference on Computational Data and Social Networks (CSoNet 2020).

and YouTube [71]. The amount of data resulting from the surge in using social media during the pandemic is vital for researchers and data scientists, which can be reflected in their behavior and perception. [72, 77].

In the past, several studies have focused on understanding people's reactions and perception of a variety of topics discussed on social media [4, 3]. Such studies focused on understanding the high volume of engagement with certain issues, and people's perception using sentiment analysis. This allows for a better understanding of users' behavior and their reactions toward a particular topic.

However, understanding and monitoring people's reactions and perception of a variety of topics discussed on social media is crucial for health officials during this pandemic [19, 47]. For instance, when the pandemic initially hits Italy, the people of Lombardy, the fourth-largest region of Italy, were sharing and discussing the possibility of a lock-down based on a CNN article [75]. Therefore, a massive crowd of people was trying to flee the region to avoid being locked-down. This, in turn, resulted in crowded public transportation and airports, which has made the overall health situation even worse. Such incidents, and other similar ones, can be at least anticipated by health officials if they were to know what is being discussed on social media and how the people are reacting to it, in order to enforce better guidance to contain the spread of the virus.

In this work, we provide an in-depth analysis of how the pandemic has affected users' interaction on Twitter, including the discussed topic, trending hashtags, mentioned users, people's perception of their sentiment, and the emojis they use. Particularly, we collected 103 million tweets from four English-speaking countries from October 2019 until November 2020, five months before the outbreak and nine months after. We then study the overall impact of the outbreak of COVID-19 on peoples' behaviors through the topics they engage with, and their sentiment toward them. We also conducted a country-based analysis to observe the different concerns and what might cause such change in the topics discussed on Twitter.

Our analysis shows that people are mostly worried about their health and economic situation. We observed such concern in the high volume of engagement during the pandemic on tweets related

to health services and the economy, where we observed an increase of 99.51% and 19.60% from February 2020 to March 2020, respectively. The analysis also showed that, when the traffic hashtag was trending in the United States, which might indicate that more people are going out, there was a spike in the cases which can be predicted from such study and prevented in the future. Moreover, we show that during such time, government officials are expected to play a huge role in addressing these issues, as in all four countries, government officials were from the most mentioned users on Twitter during the time of the pandemic.

Contributions. This work studies the social impact of user's behavior on Twitter before and during the pandemic. In particular, we make the following contributions:

- We collect a large-scale dataset of 103 million tweets from four English-speaking countries, namely, the United States, Canada, England, and Australia, collected over 14 months.
- We provide an in-depth analysis of the collected tweets, showing the shift of the topics discussed before and during the outbreak of COVID-19.
- We provide insights on who and what could be of interest to the people during the pandemic, by extracting and analyzing hashtags, mentions, and used emojis. Our analysis shows that most mentioned accounts were accounts belongs to people representing their governments, such as presidents, ministers, and government officials.
- We utilize the advancements in deep learning and natural language processing techniques to extract and track topics discussed in Twitter, along with the trends of users' behaviors towards these topics. Our analysis shows a high increase in the negative sentiments of tweets posted since COVID-19 was declared as a global pandemic. Among all categories, tweets related to "Health services" and "Jobs/Businesses" have the highest ratio of negativity in March 2020, as people were concerned about their health and losing their jobs.

Understanding and monitoring people's concerns and needs can help authority health officials to instantiate better guidance and measures to contain the spread of the virus. Therefore, in this study,

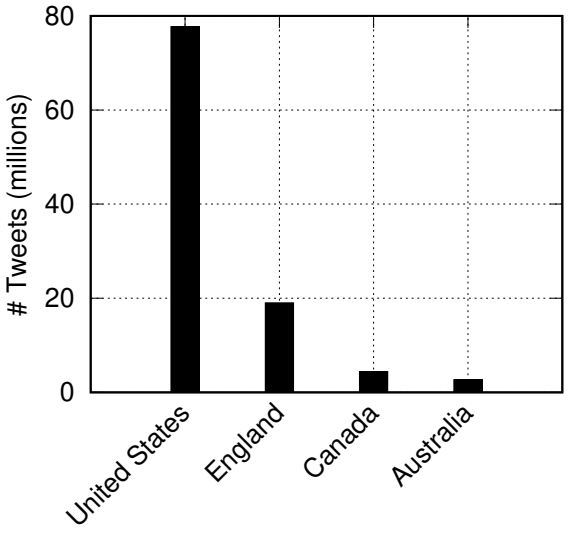


Figure 5.1: The distribution of collected tweets over countries. 74.85% of the tweets are collected from the United States.

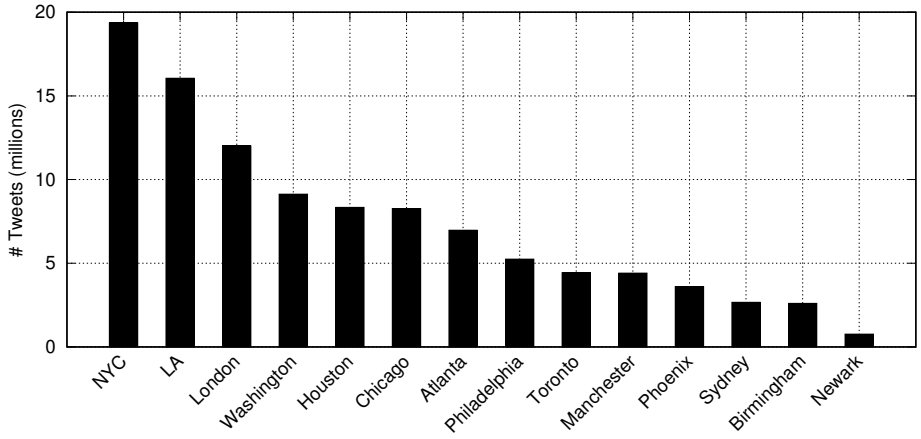


Figure 5.2: The distribution of collected tweets over different cities. Eight cities (57%) are within the United States.

we provide in-depth analysis and insights into how the pandemic has impacted people's behavior towards topics discussed on Twitter. Moreover, we analyze trends and shifts of emotions and feelings observed when discussing these topics before and after the pandemic. To do so, we utilize state-of-the-art techniques to detect, model, and track different topics from tweets collected in time-frame covering periods before and after the COVID-19 outbreak. After observing the major topics discussed during the data collection period, we adopted a deep learning-based sentiment

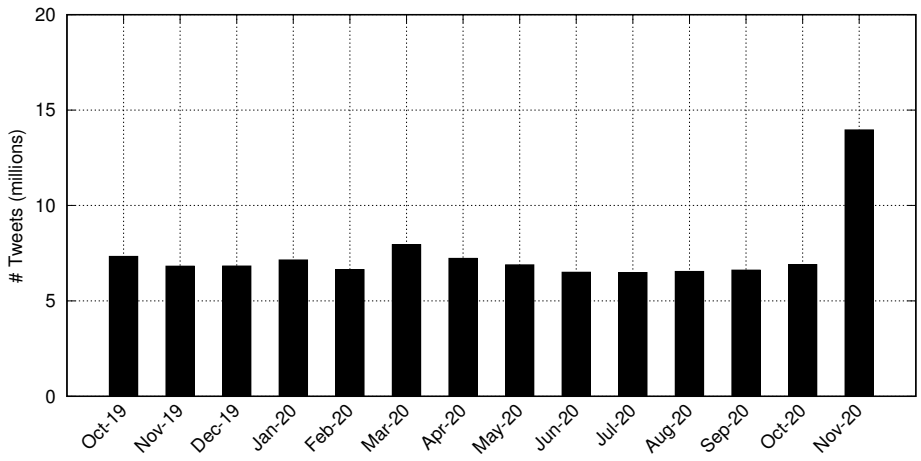


Figure 5.3: The number of collected tweets per month within the studied duration. Note that the number of tweets is evenly distributed over the months, with an average of 7 million tweets per month.

analysis to study the people's perceptions of these topics before and after the pandemic.

Data Collection and Representation

In this work, we study and analyze the shift and change in users' sentiment and behavior toward various topics before and during the outbreak of COVID-19. To this end, we collect a large-scale dataset of more than 103 million tweets from four English-speaking countries, namely: the United States, England, Canada, and Australia. The data collection process includes scraping tweets originated from cities with the highest amount of daily-tweets in the four selected countries over 14 months, starting from early October 2019 until late November 2020.

We used Twint [28], a tool for scraping tweets from Twitter profiles through the browser without directly accessing the official Twitter's API, to scrap the tweets by location and time. Twint facilitates the large and unrestricted collection of data for our analysis since the official API has a limited number of requests daily. Besides the collected tweets, we use Sentiment140 dataset [33], a collection of 1.6 million tweets, 800,000 tweets labeled as positive (*i.e.* positive sentiment), and 800,000 tweets labeled as negative (*i.e.* negative sentiment). The latter dataset is used as ground-truth for

training and evaluating the sentiment classifier.

Data Collection: Statistics and Measurements. The collected tweets are distributed over a period of 14 months across four countries, as shown in Fig. 5.1. Since the United States has by far the largest number of users on Twitter (62.55 million users) [30], 74.85% of the tweets are from the United States. Note that both Australia and Canada have fewer users compared to the United States. Therefore, we limited the data collection from these countries to one major city from each country, with the highest number of tweets originated from Toronto (Canada) and Sydney (Australia), as shown in Fig. 5.2. The number of tweets monthly has an average of 7 million tweets per month, with a peak of 14 million tweets in November of 2020 as shown in Fig. 5.3. This increase is mainly contributed to the high volume of tweets related to the 2020 United States election [76].

Data Preprocessing

In order to understand users' behavior on Twitter before and during the COVID-19 outbreak, we perform two tasks: topic modeling and sentiment analysis. The topic modeling task aims to detect topics that were discussed during the studied period while keeping track of the emerging trends related to these topics. The sentiment analysis task aims to explore people's perceptions and the reaction of topics and whether the behaviors towards these topics are impacted by the pandemic, in a correlation-based analysis.

Each task requires different data preprocessing steps. For topic modeling, we cleaned the collected tweets by removing special characters, URLs, as well as non-English characters. Then, we removed stopwords and short phrases with less than three characters. This preprocessing step is important considering the targeted task, *i.e.* topic modeling since topics are informed by keywords that occur across a large number of documents.

For the sentiment analysis task, we kept the original tweets and applied WordPiece tokenization [63] on the collected tweets. Skipping other preprocessing steps, such as eliminating stopwords and non-English characters representing emojis, improves the performance without im-

pacting the accuracy of the analysis. This is due to the employment of Bidirectional Encoder Representations from Transformers (BERT), which utilizes WordPiece tokenization in its core. WordPiece tokenization is a word segmentation algorithm that forms a new subword of a token using a pre-trained likelihood probability model, *e.g.* the word "working" is divided into two subwords; "work" and "ing". This is particularly beneficial when handling out-of-vocabulary words. We describe the use of WordPiece in the following section.

Data Representation

The topic modeling and sentiment analysis tasks require transforming the text to numerical encoding and count-based representations. For that, we used bag-of-words representation for topic modeling and WordPiece tokenizer for sentiment analysis.

Bag-of-Words Representations. For the topic modeling task, representing the documents as clusters of different topics, based on the similarity score of their context, goal, or interest, requires considering frequent terms from the content of documents. Therefore, a bag-of-words model is a well-fit candidate for this task. Bag-of-words or (BoW) for short, is a technique that extracts features from text to be used in modeling, such as machine learning algorithms. The approach is quite simple and easy, and can be used in a variety of ways for extracting features from documents. In a document, BoW is a numerical representation of a given text that calculates the occurrence of words within a document. It consists of two tasks building a dictionary of all terms from a collection of documents. Then, measuring the presence of these words for each document.

In order to utilize different state-of-the-art topic modeling techniques, tweets are first represented using a BoW during the data representation phase, in which a tweet is transformed into a vector of values that represent the presence or the weight of all unique words in the corpus in relation to the tweet. A tweet is commonly represented with a hot-encoding vector that highlights the existing words of the tweet against all terms in the dictionary. Considering the fact that we are handling over 100 million tweets that have a huge number of unique terms, the dictionary, defined

as the collection of all unique terms in a corpus, can be very large. This will produce sparse and high-dimensional BoW representations. To optimize the bag-of-words representation of tweets, we remove the most frequent and rare words. Common words that appear in more than 50% of all tweets are considered general terms with less discriminative meaning than other less frequent words. Rare words that appear less than 1,000 times in the entire corpus are also eliminated, as they only introduce noise to the data representation. We note that we create a pre-defined list of words related to COVID-19 in the final feature set, including terms such as: *coronavirus*, *virus*, *corona*, *COVID19*, *COVID-19*, and *COVID*.

Finally, and in order to reduce the dimensionality of the bag-of-words representation, we selected the most frequent 10,000 words to be the baseline features for representing the tweets.

WordPiece-based Representations. Due to the difference in the two tasks, we utilized different methods in the data representation. For the sentiment analysis task, we use a WordPiece-based representation method to represent tweets to the BERT model [16]. The WordPiece technique is based on a subword tokenization scheme with a vocabulary of size 30,000 tokens. The generated vocabulary of a fixed number of words, sub-words, and characters is helpful in addressing out-of-vocabulary words by splitting unrecognized words into sub-words, if no sub-word matches in the pre-defined dictionary, it is then split further into characters and then mapped to the corresponding embedding. This technique has been proven efficient as opposed to other embedding mechanisms that map all of the out-of-vocabulary words to one token, such as 'UNK' [63, 86].

Tweets are then represented as a matrix, where the i^{th} row contains the embedding of the i^{th} word in the corresponding tweet. For our implementation, we used 70 words as the length of tweets and 768 as the length of the word embedding vector, *i.e.* each tweet is represented as 70×768 matrix. We note that, in the collected dataset, 99.8% of the tweets were within this length. While word embedding of size 768 is the pre-defined size in the BERT model.

Investigating the trends of these topics and the impact of the pandemic on users' reactions to these topics (in a correlation analysis), requires powerful tools and techniques such as topic modeling.

LDA Topic Modeling. Topic modeling is an unsupervised machine learning technique that is typically used to extract a set of topics in a group of documents. Topic modeling processes a set of documents and detects the repeated patterns of words and phrases across documents, to cluster the documents based on their similarities [36, 9]. Using topic modeling, topics are represented with a cluster of similar and related words [10]. This enables the detection and tracking of topics through the data collection period. Accurate detection of topics allows real-time analysis and observation of trends, *e.g.* users' reactions and behavior towards topics.

This study utilizes the MALLET's Latent Dirichlet Allocation (LDA)-based topic modeling technique [42], a state-of-the-art topic modeling approach that maps each document in the corpus to a set of topics. LDA represents documents as random mixtures over latent topics, where each of the produced topics is constructed by a distribution of words. Each word w_t in a corpus $\boldsymbol{w}$ is presumably generated from a latent topic z_t, which is originated from a document-specific over set of $\boldsymbol{T}$ topics.

The word produced by the LDA is calculated by a conditional distribution $P\left(w_t = i \mid z_t = k\right)$, denoted as $T(W-1)$ free parameters, where $\boldsymbol{T}$ represents the desired number of topics and $\boldsymbol{W}$ is the size of the entire vocabulary in the given documents. The provided parameters are described by Φ, with $P\left(w_t = i \mid z_t = k\right) \equiv \phi_{i|k}$. Φ can be considered as a probability matrix, in which the k^{th} row, is the distribution of words for a given topic $\boldsymbol{k}$, is described as ϕ_k. Following the same scheme, topic generated by the LDA is defined by a conditional distribution $P\left(z_t = k \mid d_t = d\right)$, denoted by $D(T-1)$ free parameters, where $\boldsymbol{D}$ represents the number of documents in the entire collection. The aforementioned parameters construct a matrix Θ, with $P\left(z_t = k \mid d_t = d\right) \equiv \theta_{k|d}$. The d^{th} row of this matrix is the distribution of topics generated for a given document $\boldsymbol{d}$, denoted

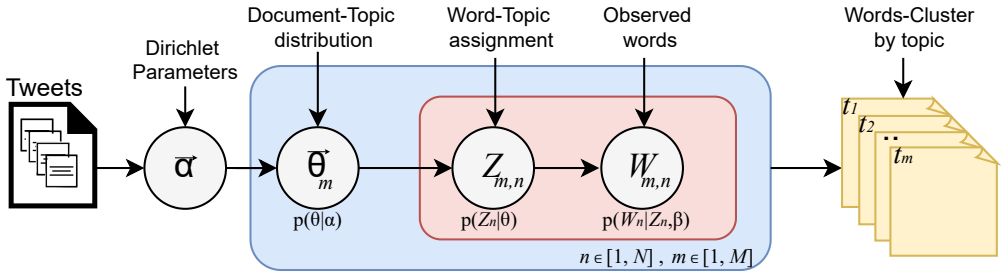


Figure 5.4: This pipeline shows the flow of for LDA topic modeling training.

by $\boldsymbol{\theta}_d$. The joint probability of a corpus $\boldsymbol{w}$ and a set of the generated latent topics $\boldsymbol{z}$ is

$$P(\boldsymbol{w}, \boldsymbol{z} \mid \Phi, \Theta) = \prod_i \prod_k \prod_d \phi_{i|k}^{N_{i|k}} \theta_{k|d}^{N_{k|d}} \tag{5.1}$$

where $N_{i|k}$ represents how many times that the word i was generated by the topic k, and $N_{k|d}$ represents the number of times that the topic k occurred in the document d.

When applying LDA, each tweet is assigned to a topic with a probability score, allowing the tweet to be recognized in different topics. For our analysis, we assign the tweet to the topic with the highest probability. The pipeline of the implemented LDA process is presented in Fig. 5.4.

LDA Configuration and Evaluation

After the data collection and preprocessing, tweets are represented using bag-of-words vector representations. Considering vectors of 10,000 bag-of-words representation as an input, the LDA model assigns topics for each tweet.

Establishing the topic model requires accurate data preprocessing and model training. In which, a number of topics are investigated in terms of the coherence score of topics. The coherence score is a score calculated for each topic by measuring the semantic similarity between words that have the highest score within the given topic. The word score is calculated based on the frequency of

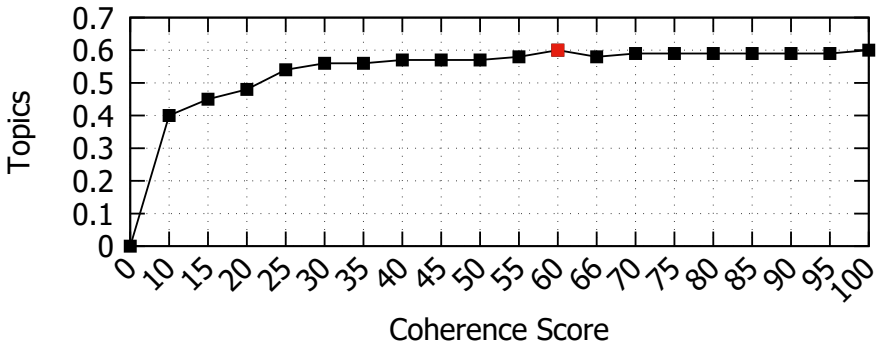


Figure 5.5: The LDA coherence score using a different number of topics.

the word within the topic and its inverse frequency with other topics as follows:

$$\text{Coherence Score } = \sum_{i<j} \text{score}\left(w_i, w_j\right)$$

where w_i, w_j are the top words of the topic. This method provides distinguishable measurements between topics that are semantically similar. The higher the coherence value, the better the quality of the clustering, indicating better topic modeling and assignment.

We examined the effect of changing the number of the extracted topics on the modeling task. In particular, we explore extracting 10–100 topics with an increase of 5 topics, while observing the coherence score achieved in each iteration. Fig. 5.5 shows the coherence score for various numbers of topics. The highest coherence score achieved was when the number of topics is 60 with a coherence score of 0.55. After inspecting the topics generated by the best model with the highest coherence score, we defined 30 distinct topics. We note that, after manually examining generated clusters of words each representing a topic, we then merged similar ones which have common keywords into one cluster. This resulted in 30 unique clusters of words out of 60 clusters that were originally produced by the model. The merging process was manually conducted based on the observations of shared common keywords and the discussed ideas.

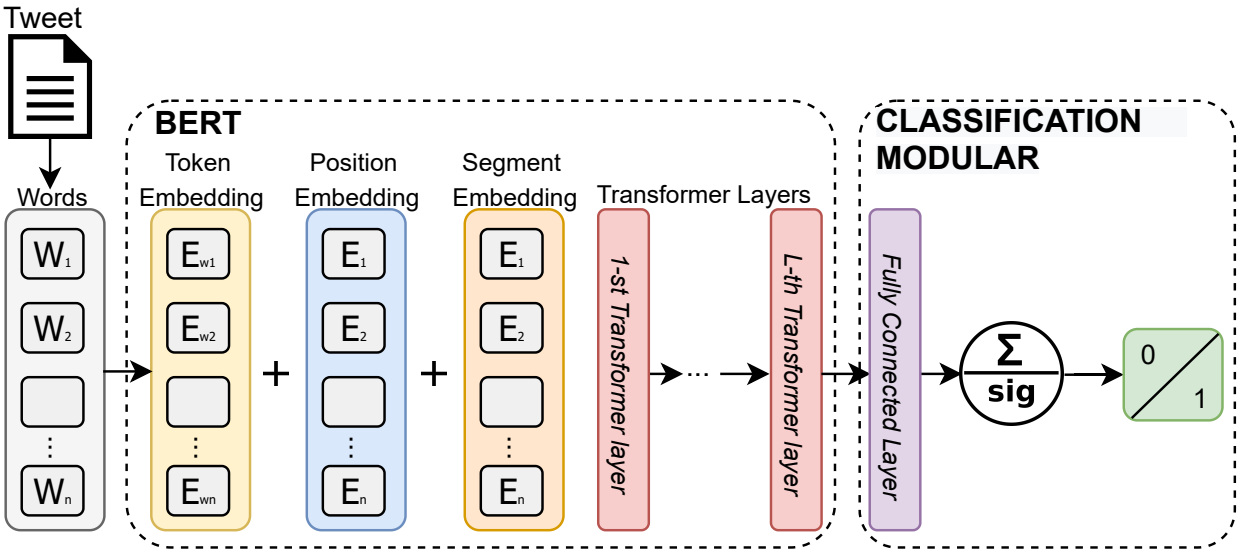


Figure 5.6: The general flow of the sentiment analysis pipeline. Sentiment140 dataset is used to fine-tune the BERT-based model for the sentiment classification task.

BERT Sentiment Analysis

The second task we aim to address is investigating the trends and shifts in perceptions of certain topics discussed on Twitter during the data collection period. This task is done by examining the sentiments observed on different topics before and during the pandemic Using Bidirectional Encoder Representations from Transformers (BERT) [16]. In its essence, BERT is a language model that benefits from the attention mechanism used in the transformer architecture[80]. The attention mechanism has two six-layers of encoders that have the ability to learn contextual relations between words in a given text, as well as six layers of decoders that generate the needed output for a given task. As opposed to traditional NLP models that read textual data sequentially from right-to-left or left-to-right, the transformer encoder is considered bidirectional since it reads the entire given text at once, allowing the model to capture the context of each word based on its surroundings. Since BERT is a language model, it uses only the encoder part of the transformer. By adding a new layer to the core model.

BERT fits a wide variety of NLP language tasks, such as classification, question answering, as

well as named entity recognition. BERT architecture follows a fine-tuning approach that does not require a particular architecture for each NLP task. This is useful in the way that it requires minimum knowledge of the model design architecture to be utilized to solve most of the NLP tasks. Instead, it should learn such knowledge from data. BERT has two pre-trained models as follow:

BERT $_{\text{BASE}}$: the base model setting has 12 layers with 768 hidden units in addition to 12 bidirectional self-attention heads and the total number of parameters in this setting is 110M

BERT $_{\text{LARGE}}$: the large model setting has 24 layers with 1024 hidden units in 16 bidirectional self-attention heads and the total number of parameters in this model is 340M. In our implementation of BERT, we used BERT $_{\text{BASE}}$ since these settings deliver the best trade-off between the accuracy and retraining time. As we can see in this pipeline Fig. 5.6, the BERT base is a stack of 12 encoders. Each of them is a transformer block. The input has to be provided to the first encoder. The BERT encoder takes a sequence of tokens (words) as shown in the pipeline and the tokens are then processed and converted. For the classification task, the special token [CLS] is inserted at the beginning of the first sentence. Then another special token denoted as [SEP] is inserted at the end of each sentence in a given text. The separated sentences are then marked to be differentiated in the segment embeddings part. In addition, the model adds the position of each token in the sequence to get position embeddings. Finally, the sum of the three embeddings becomes the final input to the BERT Encoder.

Fine-tuning BERT for Sentiment Classification

The goal of the sentiment classification task is to classify the sentiment polarity of a piece of text *e.g.*tweet to be either positive or negative. In this study, tweets are separated per sentence, with two special tokens indicating the start and the end of the tweet. Then, each tweet is fed into the trained BERT model, providing the embedding of each word in the tweet considering its surroundings. The output of the BERT model is a sigmoid activation function for binary classification signaling the polarity of the tweet, *i.e.* either positive or negative sentiment.

Table 5.1: Evaluation of deep learning models on sentiment analysis task. The BERT-based model outperforms its counterparts, therefore, used as the baseline in our analysis.

Model	True Positive Rate	True Negative Rate	Precision	F-1 score
DNN	80.15%	80.97%	80.91%	80.53%
CNN	81.51%	82.67%	82.55%	82.03%
LSTM	81.60%	81.62%	81.71%	81.65%
BERT	86.77%	87.45%	87.35%	87.06%

Sentiment Classifier Evaluation Metrics

To establish a baseline sentiment analysis model, we used a ground-truth benchmark dataset of 1.6 million annotated tweets for sentiment analysis task [33].

The sentiment classifier performance was evaluated using four evaluation metrics, namely, *True Positive Rate* (TPR), and *True Negative Rate* (TNR), *Precision*, and *F1-score*. Precision measures how good a model is in term of identifying the positive class (P = TP/TP+FP). F1-score is calculated as (F1-score $=$ 2TP/2TP+FP+FN) where TP, FP, and FN represent True Positive, False Positive, and False Negative, respectively. The TPR measures the percentage of how many positive samples are correctly identified (TPR $=$ TP/TP+FN). The TNR measures the percentage of how many negative samples are correctly identified (TNR $=$ TN/TN+FP).

We started the work by conducting several preliminary experiments to determine the best model architecture to perform the task, including deep learning-based techniques such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM), and BERT. Using the ground-truth dataset, the achieved F-1 scores were 80.53%, 82.03%, 81.65%, and 87.06% for DNN, CNN, LSTM, and BERT, respectively. Table 5.1 shows the achieved results of different models performing the sentiment analysis in terms of true positive rate, true negative rate, precision, and the F-1 score. Note that BERT outperforms all others, achieving a precision accuracy of 87.35%. Therefore, we selected the BERT model to perform the sentiment analysis in our study.

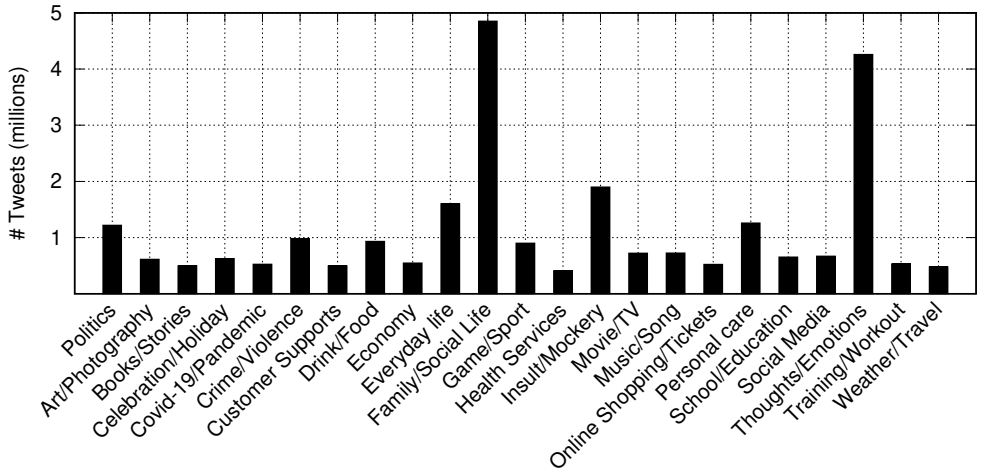


Figure 5.7: The distribution of tweets generated topics by LDA over several distinct topics.

The dataset setup used for the evaluation was split as 60% of the annotated tweets for fine-tuning the model, 20% of the data for validation, and 20% for testing the performance. By doing so, we achieved an F-1 score of 87% on the test dataset. Table 5.1 shows the evaluation of different deep learning architectures in sentiment analysis tasks on the Sentiment140 dataset. Note that the BERT-based model outperforms its counterparts; therefore, it is used as the baseline for our analysis.

Results and Discussion

Topic Modeling: Observations and Outcomes. Using the best-performing LDA model, we manually inspect the topics through the frequent keywords generated for each topic and assign names and descriptions to those topics.

Topic-Tweet Distribution. Fig. 5.7 shows the distribution of topics across the collected tweets. Clearly, due to the nature of the platform tweets related to "Family/Social Life" represent around 30% of the tweets. Also, tweets are related to "Thoughts/Emotions" have a high number of tweets with 6.2 million tweets, and tweets that have insult and mockery, around 4.9 million tweets. Even though tweets related to COVID-19 weren't discussed earlier than it was declared as a national

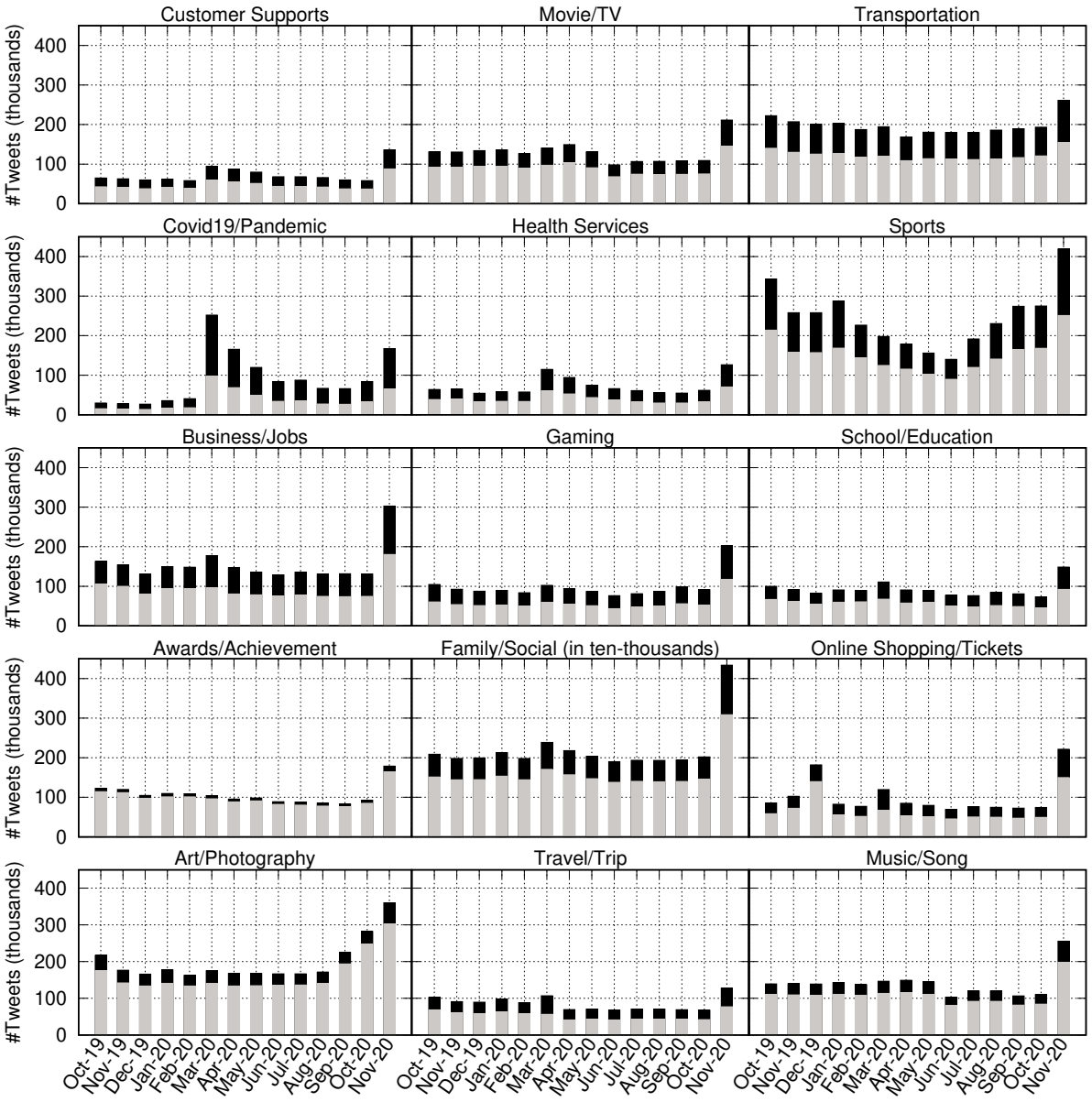


Figure 5.8: The number of both the positive and the negative tweets per generated topic from October of 2019 to November of 2020. ■ Represents the amount of the negative tweets ■ Represents the amount of the positive tweets.

pandemic, there a considerable amount of tweets of 1.3 million were COVID-19/Pandemic-related during the studied period showing the impact of the pandemic on the people's interest. Finally, most of the other topics such as Sports, Food, etc. have $approx$ 2-3 million tweets per topic.

Overall Topic Temporal Tracking. Applying the LDA model with the highest coherence score

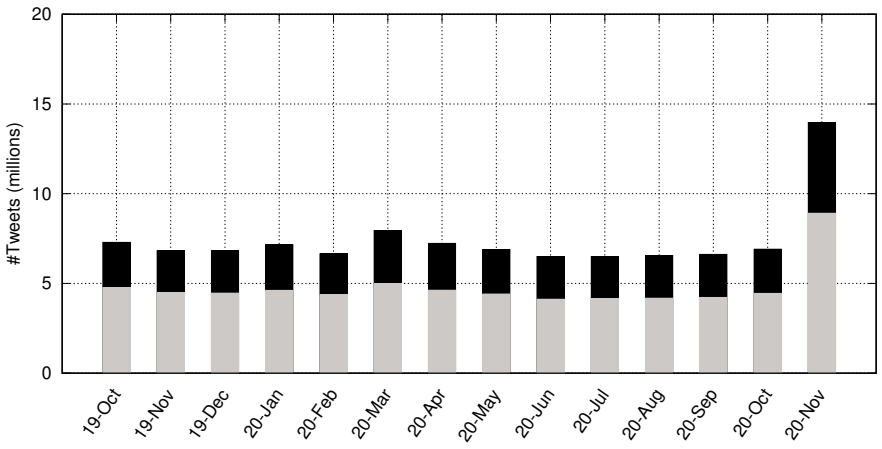


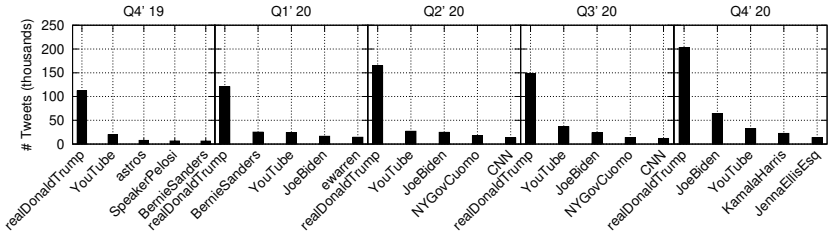
Figure 5.9: The sentiment of collected tweets throughout the pandemic. Showing the spike of the negativity in March.

to the entire dataset, we monitored tweets posted before and during the pandemic observing several trends and topic evolution. Fig. 5.8 shows the temporal distribution of tweets throughout the studied period.

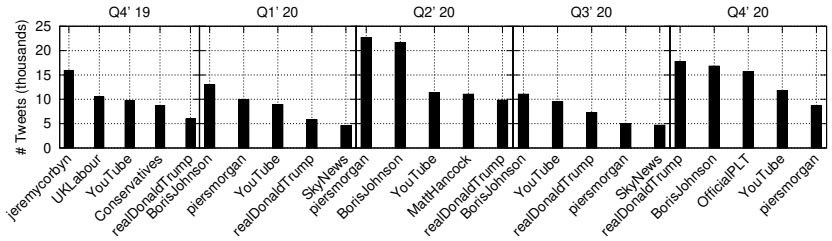
As expected, we observe a rapid increase in the tweets related to the "COVID-19/Pandemic" in March 2020, as a result of declaring COVID-19 as a global pandemic by the World Health Organization. This increase is shown in Fig. 5.8: *COVID-19*, with a 520.04% increase in the volume of the tweets from February 2020 to March 2020. This huge jump has then decreased in the next few months before increasing again at each of any new worldwide wave of the cases such as in July 2020 and October 2020. Similarly, tweets related to health services have peaked in March 2020, the same month where COVID-19 was declared a global pandemic. The tweets related to "Health Services" showed a 99.51% increase from February to March, as shown in Fig. 5.8: *Health Services*.

As the pandemic has affected many aspects of people's lives, tweets related to jobs and business have spiked, as governments and organizations started enforcing lock-downs and social distancing we had a negative impact on many jobs and businesses. Before the pandemic, tweets related to jobs and business were declining until December 2019, where they hit the lowest rate through-

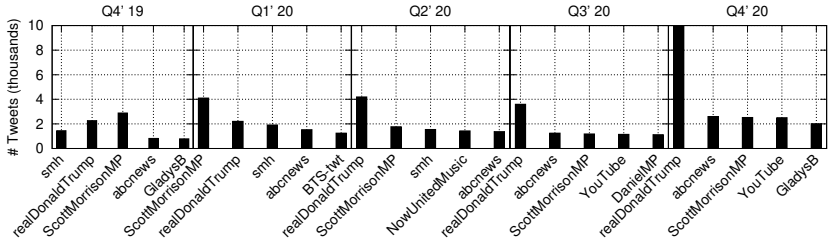
out the studied period. Then, it witnessed an increase in the number of tweets (*i.e.* 36.10%) in March 2020, as many people have lost their jobs and businesses were shutting down, as shown in Fig. 5.8: *Business/Jobs*.



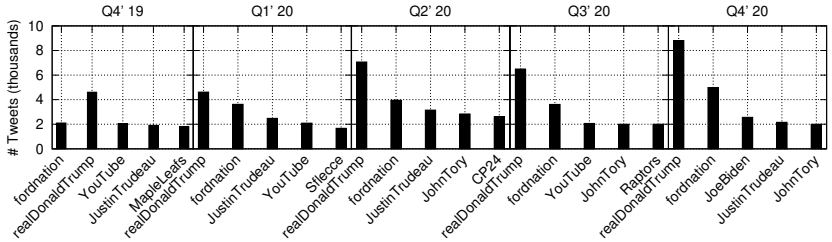
(a) USA Accounts



(b) England Accounts



(c) Australia Accounts



(d) Canada Accounts

Figure 5.10: The top five mentioned accounts in the four countries before and during the pandemic aggregated quarterly. Each country is investigated individually, starting from the last quarter of 2019 until the last quarter of 2020.

Moreover, events and activities that require large gatherings in one place, such as sports and awards

ceremonies, are highly affected due to the pandemic. This is also reflected in our analysis, as it shows a steady decline in tweets related to these topics as in Fig. 5.8: *Awards/Achievement* and Fig. 5.8: *Sports*. However, as most major sports have resumed their tournaments in July, the number of tweets related to sports has risen again even though these events have enforced some safety measures, including a limited audience.

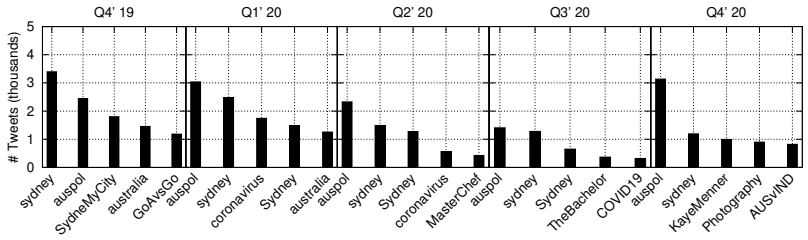
The negative impact of the pandemic on the volume of tweets is observed on topics related to travel and trip as shown in Fig. 5.8: *Travel/Trip*. The tweets decreased from 102,384 in October 2019 to 67,451 in October 2020, showing the effect of the quarantine on people's plans. On the other hand, tweets about movies/TV shows and music/songs spiked during the quarantine as shown Fig. 5.8: *Music/Song*. We then notice a decrease to less than what it was before the outbreak as shown in Fig. 5.8: *Movie/TV* and Fig. 5.8: *Music/Song*. We also observed a decrease in tweets related to "Transportation", with more than a 21% drop in volume from October 2019 to April 2020 as shown in Fig. 5.8: *Transportation*. Similar trends were observed in technology and customer support-related tweets, as people start extensively relying on technologies and using the online platform to connect and run businesses during the quarantine. We observed an increase of 45.61% of such tweets from December 2019 to March 2020.

Country-level Topic Temporal Tracking. To get a better view of the impact of the pandemic on each of the countries and how different topics correlate with the COVID-19 cases, we extracted the topic discussed for each of them to investigated if countries differ in terms of the impact of the pandemic on the social and normal life. In Australia for example, there is a clear correlation between the number of COVID-19 cases and the number of tweets related to sports. Tweets related to sports had an average of 3,500 tweets in the first three months of 2020 in which the first spike occurred and the cases had an average of $\approx$ 300 cases a day [84]. Then, in the next three months, tweets related to sports declined as the COVID-19 cases went down to an average of $\approx$ 15 cases a day. And the same pattern continues for the next spike the average number of tweets related to sports was higher. This might be an indication of a lack of health prevention methods used in stadium and

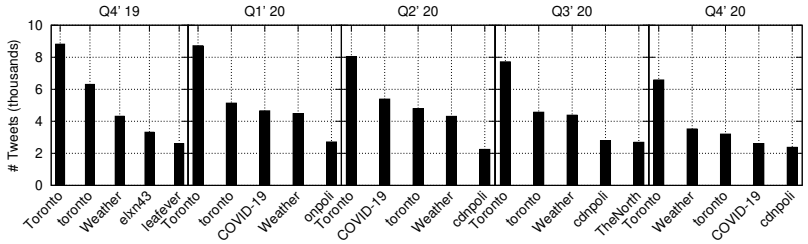
sports events. Other findings we observed which is a common concern across all countries, school and education-related tweets increased drastically in March where most governments implemented lockdown restrictions to restrict citizens' movements avoiding any unnecessary gathering of people. The effect of this restriction is also noticed with a spike in the online shopping-related tweets which has witnessed an increase of 100% over the average number of tweets except during Christmas time. The aggressive measures which include curfew for the big cities taken by the Australian government in August of 2020 [7] can be clearly seen in our analysis wherein August tweets related to trip/travel were at their lowest rate across all of the studied period.

Interestingly, in Canada during the black-lives movement In June of 2020 where tweets about protesting were very high around 7,000 where it is normally around 3,000 tweets. During such movement it is expected to see a spike in the cases since people are marching in large groups, however, the new cases decreased by 35% from the previous month. This might show the importance of people's awareness of in considering the safety measures and by keeping the recommended distances between others and wearing masks. Our analysis also showed that tweets related to trip/-travel have been declining since March of 2020 when the Canadian government imposed travel restrictions on non-essential travel. This restriction has been in effect through the remainder of 2020 [1]. One of the concerns the Canadian people raised was the fear of losing their jobs and business as tweets related to job/business were at their highest point. Such concern was due to the fact the employment in Canada fell by more than one million in March [55].

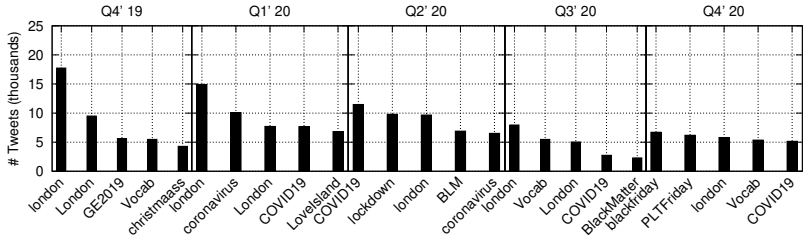
The national health service in England was overwhelmingly impacted by the pandemic, starting from primary care to hospital treatment, resulting in a lack of care for a lot of patients. As many appointments were canceled or conducted remotely to keep both health workers and patients safe [62]. Such an incident has been reflected in the perception of the people of England where tweets related to health services have spiked when the pandemic first hits England. And when the prime minister of England instructed people to stay at home [2] after a surge in coronavirus cases, there was an increase in the number of online service usage showing the importance and impact of the authorities when handling such situations. Our analysis showed that, since March where



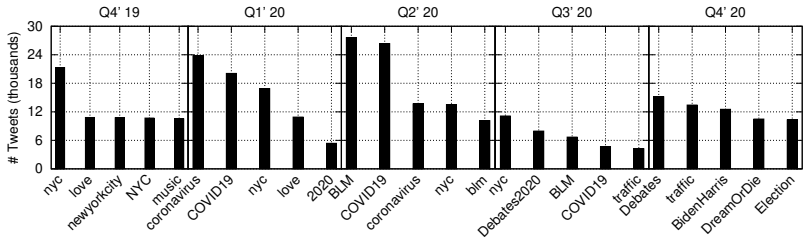
(a) Australia Hashtags



(b) Canada Hashtags



(c) England Hashtags



(d) USA Hashtags

Figure 5.11: The top five trending hashtags in the four countries before and during the pandemic aggregated quarterly. Each country is investigated individually, starting from the last quarter of 2019 until the last quarter of 2020.

the ”stay-at-home“ order was broadcasted, tweets related to online shopping as well as customer supports were increasing. On the other hand, the safety measures continued throughout the year, the tweets discussing normal and social life have been in a consistent drop since then.

The U.S. Centers for Disease Control and Prevention (CDC) activated its Emergency Operations

Center(EOC) in late January of 2020 in order to limit the spread of COVID-19 and support public health to better respond to the outbreak. Since then, some businesses have started to close and laying down employees following the health guidance. As a result of such measures people addressed their concerns in their tweets where tweets related to job/business kept increasing and peaked in March at the same time the US payrolls dropped by 701,000 [31]. Moreover, our analysis shows an increase in the tweets about gaming during the quarantine in the United States where the number of US gamers went up from 214 million in 2019 to 227 million in 2020 [43]. Meanwhile, as in other countries, online services such as online shopping and customer supports have also increased amid the pandemic.

Topics-derived Sentiment

Fig. 5.9 shows the overall positive and negative tweets over 14 months. We observed that the highest negativity rate was in tweets posted in March 2020, when COVID-19 was declared as a global pandemic. Among all categories, tweets related to "COVID-19/Pandemic", "Health services", and "Jobs/Businesses" have the highest negativity ratio in the same month, where people were concerned about their health and losing their jobs. Overall, before the pandemic, tweets related to Art/Photograp have been mostly with negative sentiments. However, after the pandemic, people have become more positive in tweets associated with Art/Photograp, as shown in Fig. 5.8: *Art/Photograp*. Interestingly, tweets related to gaming had been positive until March, as people started showing some negativity. "Health Services" related tweets showed one of the highest negativity rates during the pandemic. After that, we observed a change in the sentiments, as people become more positive as time progress, as shown in Fig. 5.8: *Health Services*, just to spike again in November 2020 which might be caused by its importance to be discussed during the US elections.

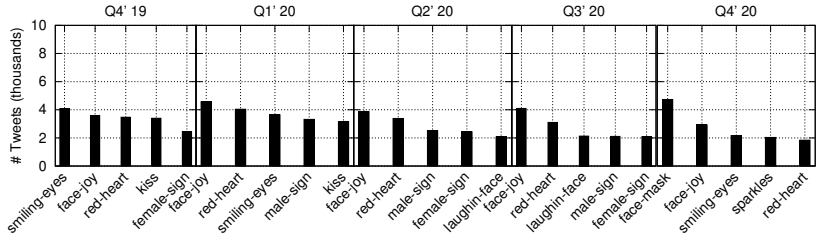
The results also show an increase in the positivity for tweets related to "Online Shopping", after the highest negativity ratio in March. This may be because more people relied on online shopping as the main alternative to regular shopping. Moreover, across all topics, school and education had

the highest negativity rate when the outbreak starts. After March, tweets were positive until the second spike, where the negativity increased again, as shown in Fig. 5.8: *School/Education*. While it is difficult to accurately pinpoint the root causes of sentiments observed on our dataset, it seems that the pandemic and associated measures are affecting people, raising their concerns, and as a result, causing noticeable trends in sentiments across various topics.

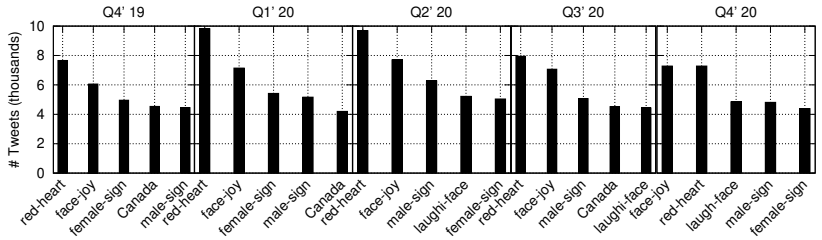
Users' Interaction

Besides analyzing the trends and sentiment of Twitter's users and understanding users' behavior on Twitter before and during the COVID-19 outbreak, we conduct a comprehensive analysis of the most mentioned accounts, used emojis, as well as trending hashtags. This provides a better understating of the shift of people's interests and feelings, as emojis can be used as an easy way of expressing people's feelings. To this end, we use regular expressions methods to extract accounts, hashtags, and emojis from all tweets. Then, we count the occurrences of each of the features aggregated quarterly. In the analysis, we investigate each country individually, starting from the last quarter of 2019 until the last quarter of 2020.

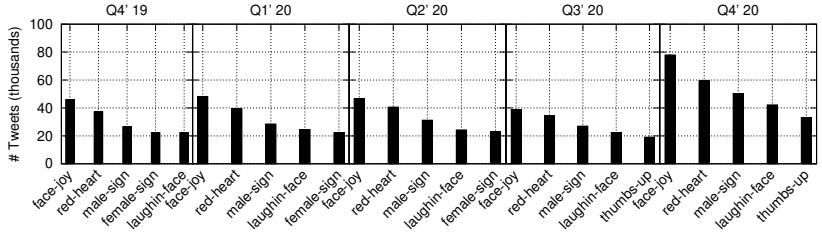
Throughout the data collection period, the president of the United States (as of November 2020) was among the most mentioned users in the four countries before and during the pandemic which might indicate the strong role that the United States has worldwide. We also observed that after the outbreak, most mentioned accounts in all countries were accounts of people representing their governments, such as ministers, leaders, and other officials presumably looking for guidance and help during such a crisis. For instance, even though the prime minister of the United Kingdom was not among the most mentioned accounts before the outbreak (in the last quarter of 2019), as shown in Fig. 5.10, he was one of the most mentioned accounts in England during the pandemic. In the first quarter of 2020, he was the most mentioned user in England. Similarly, in Australia, the prime minister was the most mentioned accounts during the first quarter of 2020 when COVID-19 hits Australia.



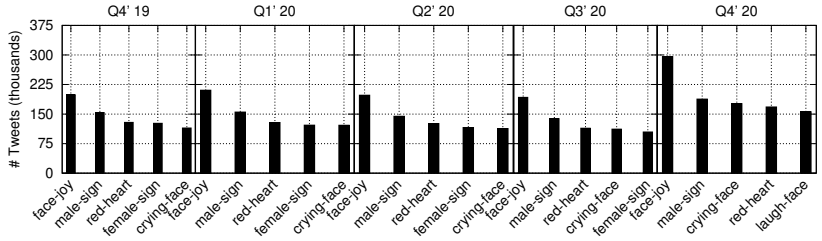
(a) Australia Emojis



(b) Canada Emojis



(c) England Emojis



(d) USA Emojis

Figure 5.12: The top five most used emojis in the four countries before and during the pandemic aggregated quarterly. Each country is investigated individually, starting from the last quarter of 2019 until the last quarter of 2020.

Further, we observe obvious trends on hashtags related to COVID-19 in all countries in the first quarter of 2020, as shown in Fig. 5.11. Such hashtags persisted during the second quarter of 2020 in countries such as England, as the number of COVID-19 cases increased [20]. We also observe that in England, the hashtag persist during the second quarter of 2020, and by the end of that

quarter the average number of daily cases went down from $\approx$ 4,000 to $\approx$ 800. Such an example might show the effect of people's awareness and mentality in encouraging others to undertake the necessary precaution. This trend is accompanied by a significant decrease in the number of cases, about 733%, in England by the end of the second quarter of 2020.

We also observed that the hashtag "traffic" was trending in the United States during the third and fourth quarter of 2020, which possibly indicates that people resume their normal lief going outside which causes traffic. During that time, the country had its third and worst wave of COVID-19 cases, where the number of cases increased by 400% from the previous quarter [20].

Studying user sentiment can be a good indication and reflection of people's perceptions of different topics. However, as many users prefer to express their feelings using emojis, we conducted an analysis of the most used emojis before and throughout the pandemic as shown in Fig. 5.12. While there was no clear change of the used emojis in Canada, Australia, and England as shown in 5.12(a), 5.12(b), and 5.12(c), respectively. The emoji of loudly crying face has become more popular in the United States during the last two quarters of 2020, as more cases of the virus were reported and the average death rate has doubled [20] as shown in 5.12(d).

Summary

In this work, we study and analyze the effect of the pandemic on Twitter's users' behaviors and sentiments, including the topic discussed, trending hashtags, mentioned users, and people's perception from their sentiment and the emojis they use. Using a large-scale dataset of 103 million English tweets before and during the pandemic from four English-speaking countries and 14 major cities, we conducted topic modeling across 30 topics and performed a temporal and semantic analysis on a monthly basis. Specifically, we study the impact of the outbreak of COVID-19 on peoples' behaviors through the topics they engage with and their sentiment toward them. Our results show that people are mostly concern with their health and economic situation. We noticed such concern by observing the high volume of engagement during the pandemic with tweets related to health

services, which witnessed an increase of 99.51% from February to March, and the tweets related to jobs and businesses, which witnessed an increase of 19.60% during the same period. Moreover, our findings show that during the pandemic, people tend to call out government officials to act and respond to the rising challenges and issues, as government officials were the most mentioned users on Twitter during the studied period. This study highlights the impotence of understanding what is people's cancers on social media, to help authorities and health officials to make better and more informed decisions when handling such pandemics.

CHAPTER 6: CONCLUSION AND FUTURE WORK

Social media platforms have been growing at a rapid pace, attracting users' engagement with the online content due to their convenience facilitated by many useful features. The growth in social media data is becoming a vital new area for scholars and researchers to explore the public's behavior and opinion pursuing different venues in social media research.

In this dissertation, we focused on studying, detecting and analyzing users' exposure to different types of toxicity on different social media platforms utilizing state-of-art techniques in both deep learning and natural language processing areas, and facilitated by exclusively collected and datasets that address various issues. The different issues, or applications, benefit from a unified and versatile pipeline that could be applied to various scenarios. The issues we studied include: (1) the detection and measurement of kids' exposure to inappropriate comments posted on YouTube videos targeting young users, (2) the association between topics of contents cover by mainstream news media and the toxicity of the comments and interactions by users, (3) the user interaction with, sentiment, and general behavior towards different topics discussed in social media platforms in light of major events (i.e., the outbreak of the COVID-19 pandemic).

The applications and the social issues we investigated in this dissertation open up future research directions, including the following

- In kids' online safety, it would be interesting to extend the study of toxicity to the gaming space where players from different age groups engage in toxic discussion. Detecting and Identifying such users is vital work that can be pursued in the future to ensure the safety of children on those platforms.
- Our study highlighted a massive amount of toxic comments posted on YouTube videos regardless of the age group, and that level of toxicity has been steady throughout many years with no noticeable changes. Such findings urge exploring the causes of such behavior as well as providing reliable and robust moderation techniques for a future direction.

- For future work in studying users' behavior and perception in online discussion, it would be interesting to pinpoint the cause root of or the start that led to change of the subject or caused users to react negatively. This direction help in both understanding the change in users' behavior as well as comment moderation.

APPENDIX: COPYRIGHT INFORMATION

Consent to Publish

Lecture Notes in Computer Science

Title of the Book or Conference Name: The 9th International Conference on Computational Data & Social Networks

Volume Editor(s) Name(s): Sriram Chellappan, Kim-Kwang Raymond Choo and Hai Phan

Title of the Contribution: An Analysis of Users Engagement on Twitter During the COVID-19 Pandemic: Topical Trends and Sentiments

Author(s) Full Name(s): Sultan Alshamari, Ahmed Abusnaina, Mohammed Abuhamad, Anho Lee, DaeHun Nyang, David Mohaisen

Corresponding Author's Name, Affiliation Address, and Email:

Sultan Alshamari, University of Central Florida, Orlando, FL 32816, USA, salshamrani@knights.ucf.edu

When Author is more than one person the expression "Author" as used in this agreement will apply collectively unless otherwise indicated.

The Publisher intends to publish the Work under the imprint **Springer**. The Work may be published in the book series **Lecture Notes in Computer Science (LNCS, LNAI or LNBI)**.

§ 1 Rights Granted

Author hereby grants and assigns to **Springer Nature Switzerland AG, Gewerbestrasse 11, 6330 Cham, Switzerland** (hereinafter called **Publisher**) the exclusive, sole, permanent, world-wide, transferable, sub-licensable and unlimited right to reproduce, publish, distribute, transmit, make available or otherwise communicate to the public, translate, publicly perform, archive, store, lease or lend and sell the Contribution or parts thereof individually or together with other works in any language, in all revisions and versions (including soft cover, book club and collected editions, anthologies, advance printing, reprints or print to order, microfilm editions, audiograms and videograms), in all forms and media of expression including in electronic form (including offline and online use, push or pull technologies, use in databases and data networks (e.g. the Internet) for display, print and storing on any and all stationary or portable end-user devices, e.g. text readers, audio, video or interactive devices, and for use in multimedia or interactive versions as well as for the display or transmission of the Contribution or parts thereof in data networks or search engines, and posting the Contribution on social media accounts closely related to the Work), in whole, in part or in abridged form, in each case as now known or developed in the future, including the right to grant further time-limited or permanent rights. Publisher especially has the right to permit others to use individual illustrations, tables or text quotations and may use the Contribution for advertising purposes. For the purposes of use in electronic forms, Publisher may adjust the Contribution to the respective form of use and include links (e.g. frames or inline-links) or otherwise combine it with other works and/or remove links or combinations with other works provided in the Contribution. For the avoidance of doubt, all provisions of this contract apply regardless of whether the Contribution and/or the Work itself constitutes a database under applicable copyright laws or not.

The copyright in the Contribution shall be vested in the name of Publisher. Author has asserted his/her right(s) to be identified as the originator of this Contribution in all editions and versions of the Work and parts thereof, published in all forms and media. Publisher may take, either in its own name or in that of Author, any necessary steps to protect the rights granted under this Agreement against infringement by third parties. It will have a copyright notice inserted into all editions of the Work and on the Contribution according to the provisions of the Universal Copyright Convention (UCC).

The parties acknowledge that there may be no basis for claim of copyright in the United States to a Contribution prepared by an officer or employee of the United States government as part of that person's official duties. If the Contribution was performed under a United States government contract, but Author is not a United States government employee, Publisher grants the United States government royalty-free permission to reproduce all or part of the Contribution and to authorise others to do so for United States government purposes. If the Contribution was prepared or published by or under the direction or control of the Crown (i.e., the constitutional monarch of the Commonwealth realm) or any Crown government department, the copyright in the Contribution shall, subject to any

agreement with Author, belong to the Crown. If Author is an officer or employee of the United States government or of the Crown, reference will be made to this status on the signature page.

§ 2 Rights Retained by Author

Author retains, in addition to uses permitted by law, the right to communicate the content of the Contribution to other research colleagues, to share the Contribution with them in manuscript form, to perform or present the Contribution or to use the content for non-commercial internal and educational purposes, provided the original source of publication is cited according to the current citation standards in any printed or electronic materials. Author retains the right to republish the Contribution in any collection consisting solely of Author's own works without charge, subject to ensuring that the publication of the Publisher is properly credited and that the relevant copyright notice is repeated verbatim. Author may self-archive an author-created version of his/her Contribution on his/her own website and/or the repository of Author's department or faculty. Author may also deposit this version on his/her funder's or funder's designated repository at the funder's request or as a result of a legal obligation. He/she may not use the Publisher's PDF version, which is posted on the Publisher's platforms, for the purpose of self-archiving or deposit. Furthermore, Author may only post his/her own version, provided acknowledgment is given to the original source of publication and a link is inserted to the published article on the Publisher's website. The link must be provided by inserting the DOI number of the article in the following sentence: "The final authenticated version is available online at https://doi.org/[insert DOI]." The DOI (Digital Object Identifier) can be found at the bottom of the first page of the published paper.
Prior versions of the Contribution published on non-commercial pre-print servers like ArXiv/CoRR and HAL can remain on these servers and/or can be updated with Author's accepted version. The final published version (in pdf or html/xml format) cannot be used for this purpose. Acknowledgment needs to be given to the final publication and a link must be inserted to the published Contribution on the Publisher's website, by inserting the DOI number of the article in the following sentence: "The final authenticated publication is available online at https://doi.org/[insert DOI]".
Author retains the right to use his/her Contribution for his/her further scientific career by including the final published paper in his/her dissertation or doctoral thesis provided acknowledgment is given to the original source of publication. Author also retains the right to use, without having to pay a fee and without having to inform the Publisher, parts of the Contribution (e.g. illustrations) for inclusion in future work. Authors may publish an extended version of their proceedings paper as a journal article provided the following principles are adhered to: a) the extended version includes at least 30% new material, b) the original publication is cited, and c) it includes an explicit statement about the increment (e.g., new results, better description of materials, etc.).

§ 3 Warranties

Author agrees, at the request of Publisher, to execute all documents and do all things reasonably required by Publisher in order to confer to Publisher all rights intended to be granted under this Agreement. Author warrants that the Contribution is original except for such excerpts from copyrighted works (including illustrations, tables, animations and text quotations) as may be included with the permission of the copyright holder thereof, in which case(s) Author is required to obtain written permission to the extent necessary and to indicate the precise sources of the excerpts in the manuscript. Author is also requested to store the signed permission forms and to make them available to Publisher if required.
Author warrants that Author is entitled to grant the rights in accordance with Clause 1 "Rights Granted", that Author has not assigned such rights to third parties, that the Contribution has not heretofore been published in whole or in part, that the Contribution contains no libellous or defamatory statements and does not infringe on any copyright, trademark, patent, statutory right or proprietary right of others, including rights obtained through licences. Author agrees to amend the Contribution to remove any potential obscenity, defamation, libel, malicious falsehood or otherwise unlawful part(s) identified at any time. Any such removal or alteration shall not affect the warranty given by Author in this Agreement.

§ 4 Delivery of Contribution and Publication

Author agrees to deliver to the responsible Volume Editor (for conferences, usually one of the Program Chairs), on a date to be agreed upon, the manuscript created according to the Publisher's Instructions for Authors. Publisher will undertake the reproduction and distribution of the Contribution at its own expense and risk. After submission of the Consent to Publish form signed by the Corresponding Author, changes of authorship, or in the order of the authors listed, will not be accepted by the Publisher.

§ 5 Author's Discount for Books

Author is entitled to purchase for his/her personal use (if ordered directly from Publisher) the Work or other books published by Publisher at a discount of 40% off the list price for as long as there is a contractual arrangement between Author and Publisher and subject to applicable book price regulation.
Resale of such copies is not permitted.

§ 6 Governing Law and Jurisdiction

If any difference shall arise between Author and Publisher concerning the meaning of this Agreement or the rights and liabilities of the parties, the parties shall engage in good faith discussions to attempt to seek a mutually satisfactory resolution of the dispute. This agreement shall be governed by, and shall be construed in accordance with, the laws of Switzerland. The courts of Zug, Switzerland shall have the exclusive jurisdiction.

Corresponding Author signs for and accepts responsibility for releasing this material on behalf of any and all Co-Authors.

Signature of Corresponding Author: **Date:**

sultan alshamrani 10/14/2020

☐ I'm an employee of the US Government and transfer the rights to the extent transferable (Title 17 §105 U.S.C. applies)

☐ I'm an employee of the Crown and copyright on the Contribution belongs to the Crown

For internal use only:
Legal Entity Number: 1128 Springer Nature Switzerland AG
Springer-C-CTP-07/2018

IEEE COPYRIGHT AND CONSENT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

Hiding in Plain Sight: A Measurement and Analysis of Kids\u2019 Exposure to Malicious URLs on YouTube
Sultan Alshamrani, Ahmed Abusnaina and David Mohaisen
2020 IEEE/ACM Symposium on Edge Computing (SEC)

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.
2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the IEEE PSPB Operations Manual.
4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.
5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

You have indicated that you DO wish to have video/audio recordings made of your conference presentation under terms and conditions set forth in "Consent and Release."

CONSENT AND RELEASE

1. In the event the author makes a presentation based upon the Work at a conference hosted or sponsored in whole or in part by the IEEE, the author, in consideration for his/her participation in the conference, hereby grants the IEEE the unlimited, worldwide, irrevocable permission to use, distribute, publish, license, exhibit, record, digitize, broadcast, reproduce and archive, in any format or medium, whether now known or hereafter developed: (a) his/her presentation and comments at the conference; (b) any written materials or multimedia files used in connection with his/her presentation; and (c) any recorded interviews of him/her (collectively, the "Presentation"). The permission granted includes the transcription and reproduction of the Presentation for inclusion in products sold or distributed by IEEE and live or recorded broadcast of the Presentation during or after the conference.
2. In connection with the permission granted in Section 1, the author hereby grants IEEE the unlimited, worldwide, irrevocable right to use his/her name, picture, likeness, voice and biographical information as part of the advertisement, distribution and sale of products incorporating the Work or Presentation, and releases IEEE from any claim based on right of privacy or publicity.

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Sultan Alshamrani	24-09-2020
Signature	**Date (dd-mm-yyyy)**

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the

IEEE work with a Digital Object Identifier (DOI) or link to the article abstract in IEEE Xplore, or (2) the accepted version only (not the IEEE-published version), including the IEEE copyright notice and full citation, with a link to the final, published article in IEEE Xplore.

Questions about the submission of the form or manuscript must be sent to the publication's editor.
Please direct all questions about IEEE copyright policy to:
IEEE Intellectual Property Rights Office, copyrights@ieee.org, +1-732-562-3966

CEUR Workshop Proceedings http://ceur-ws.org ISSN 1613-0073

Version 1.03-ntp
2020-03-02

CEUR Workshop Proceedings
http://CEUR-WS.org
RWTH Aachen, Informatik 5, Aachen, Germany

Author Agreement to Publish a Contribution as Open-Access on CEUR-WS.org

Herewith I/we (the author(s) resp. the copyright holders) **agree** that my/our contribution:

authored by: Sultan Alshamrani, Mohammed Abuhamad , Ahmed Abusnaina ,and David Mohaisen.

with corresponding author

Name: Sultan Alshamrani

Affiliation: University of Central Florida

Address: 4000 Central Florida Blvd, Orlando, FL 32816

Email: salshamrani@knights.ucf.edu

shall be made available as an open-access publication under the **Creative Commons License Attribution 4.0 International (CC BY 4.0)**, available at https://creativecommons.org/licenses/by/4.0/legalcode, and be published as part of the proceedings volume of the event

Name and year of the event: MAISoN 2020

Editors of the proceedings (editors): Ebrahim Bagheri, Huan Liu, Kai Shu, Fattane Zarrinkalam

I/we agree that my/our contribution is made available publicly under the aforementioned license on the servers of CEUR Workshop Proceedings (CEUR-WS). I/we grant the editors, RWTH Aachen, CEUR-WS, and its archiving partners the non-exclusive and irrevocable **right to archive** my/our contribution and **to make it accessible** (online and free of charge) for **public distribution**. This granted right extends to any associated metadata of my/our contribution. Specifically, I/we license the associated metadata under a Creative Commons CC0 1.0 Universal license (public domain). I/we agree that our author names and affiliations is part of the associated metadata and may be stored on the servers of CEUR-WS and made available under the CC0 license. I/we acknowledge that the editors hold the copyright for the proceedings volume of the aforementioned event as the official collection of contributions to the event.

I/we have not included any copyrighted third-party material such as figures, code, data sets and others in the contribution to be published.

I/we warrant that my/our contribution (including any accompanying material such as data sets) does not infringe any rights of third parties, for example trademark rights, privacy rights, and intellectual property rights. I/ we understand that I/we retain the copyright to my/our contribution. I/we understand that the dedication of my/our contribution under the CC BY 4.0 license is irrevocable.
I/we understand and agree that the **full responsibility/liability** for the content of the contribution rests upon me/us as the authors of the contribution. I/we release the aforementioned editors, RWTH Aachen, persons providing the CEUR-WS service, and the archiving partners of CEUR-WS from any liability caused by the publication or archiving of my/our contribution via the servers used by CEUR-WS.
I/we have read the conditions of the Creative Commons License Attribution 4.0 International (CC BY 4.0), and agree to apply this license to my/our contribution.

Location, Date, Signature of the corresponding author representing all authors
(Signature must be handwritten with a pen on paper)

IW3C2 Copyright Release Form

Title of work:	**Hate, Obscenity, and Insults: Measuring the Exposure of Children to Inappropriate Comments in YouTube**
Author(s):	Sultan Alshamrani, University of Central Florida, Ahmed Abusnaina, University of Central Florida, Mohammed Abuhamad, Loyola University Chicago, Daehun Nyang, Ewha Womans University, David Mohaisen, University of Central Florida
Description of material:	**WWW Paper**
Title of IW3C2 Publication:	**WWW '21: The Web Conference 2021**

I hereby assign (to the extend transferable, see point B below) to the International World Wide Web Conference Committee (IW3C2) the copyright of this Work for the full period of copyright and all renewals, extensions, revisions and revivals together with all accrued rights of action throughout the world in any form, including as part of IW3C2 and the public Conference Web site, on CD-ROM and in translation, or on videocassette, broadcast, cablecast, laserdisc, multimedia or any other media format now or hereafter known. (Not all forms of media will be utilized.) I accept that IW3C2 will allow the Association for Computing Machinery (ACM) to distribute, make available on the Internet or sell this Material as part of the above-named publication in the ACM Digital Library. Finally, I also accept that IW3C2 will publish this Work under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license, which reserves my rights to disseminate the work on my personal and corporate Web site with the appropriate attribution. Notwithstanding the above, I retain all proprietary rights other than copyright as assigned above, such as patent and trademark rights. For further details, see the Appendix below.

In the event that any elements used in the Material contain the work of third-party individuals, I understand that it is my responsibility to secure any necessary permissions and/or licenses and will provide it in writing to IW3C2. If the copyright holder requires a citation to a copyrighted work, I have obtained the correct wording and have included it in the designated space in the text.

I hereby release and discharge IW3C2 and other publication sponsors and organizers from any and all liability arising out of my inclusion in the publication, or in connection with the performance of any of the activities described in this document as permitted herein. This includes, but is not limited to, my right of privacy or publicity, copyright, patent rights, trade secret rights, moral rights, or trademark rights.

All permissions and releases granted by me herein shall be effective in perpetuity and throughout the universe, and extend and apply to the IW3C2 and its assigns, contractors, sub-licensed distributors, successors, and agents.

The following statement of copyright ownership will be displayed with the Material, unless otherwise specified: **" © [year] International World Wide Web Conference Committee (IW3C2), published under Creative Commons CC-BY 4.0 License." IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.**

(●) A. I hereby represent and warrant that I am the sole owner (or authorized agent of the copyright owner(s)) and assign publishing rights

( ) B. I do not own some rights to this work: (check applicable)

Audio/Video Release

* Your Audio/Video Release is conditional upon you agreeing to the terms set out below.

I further grant permission for ACM to record and/or transcribe and reproduce my presentation and likeness in the conference publication and as part of the ACM Digital Library and to distribute the same for sale in complete or partial form as part of an ACM product on CD-ROM, DVD, webcast, USB device, streaming video or any other media format now or hereafter known. I understand that my presentation will not be sold separately as a stand-alone product without my direct consent.

Accordingly, I further grant permission for ACM to include my name, likeness,

presentation and comments and any biographical material submitted by me in connection with the conference and/or publication, whether used in excerpts or in full, for distribution described above and for any associated advertising or exhibition.

Do you agree to the recording, transcription and distribution? ◉ Yes ○ No

Assign Rights

☑ I hereby assign rights and agree to publish under Creative Commons.

Third Party Material

☐ I have used third-party material and have permission to do so.

DATE: **02/23/2021** - salshamrani@knights.ucf.edu

Appendix

For details on the license used to publish the material by IW3C2, and the rights that are thereby granted, please consult the description of the Creative Commons Attribution 4.0 International (CC-BY 4.0) license on the Web Site of Creative Commons: https://creativecommons.org/licenses/by/4.0/. In case of republication, reuse, etc., the following attribution should be used: "Published in [include the complete citation information for the final version of the Work as published in the ACM edition of the Proceedings] © [year] International World Wide Web Conference Committee, published under Creative Commons CC-BY 4.0 License."

Applicable Law

The law governing this Agreement is the law of Switzerland. Any dispute concerning this Agreement shall be subject to the non-exclusive jurisdiction of the courts of Switzerland.

In addition, the following rights statement of copyright ownership needs to be displayed with the Material (on the first page of the pdf), unless otherwise specified:

For your reference the DOI assigned to your submission is:
https://doi.org/10.1145/3442442.3452314

NOTE: The DOIs will be registered and become active shortly after publication in the ACM Digital Library (closer to the conference date).

Please be sure you have followed the preparation instructions as provided by Sheridan Communications at: https://www.scomminc.com/pp/acmsig/www2021.htm which utilizes the appropriate new TAPS WORD template and/or the new ACM Consolidated LATEX Template (Version 1.57a). In order to generate accessibility friendly responsive HTML5 and PDF output files, ACM requires you download and use the new ACM Word or LaTeX template per the Sheridan Communications instructions.

Please copy and paste \setcopyright{iw3c2w3} before \begin{document} and Please copy and paste the following code snippet into your TeX file between \begin{document} and \maketitle, either after or before CCS codes.

LIST OF REFERENCES

[1] Stephanie Tam, Shivani Sood and Chris Johnston. Impact of COVID-19 on the tourism sector, second quarter of 2021, 2020.

[2] Adam Bienkov , Adam Payne , and Thomas Colson . The UK has gone into full coronavirus lockdown with the public barred from leaving home for nonessential reasons, 2020.

[3] S. Alshamrani, M. Abuhamad, A. Abusnaina, and D. Mohaisen. Investigating online toxicity in users interactions with the mainstream media channels on youtube. In *The 5th International Workshop on Mining Actionable Insights from Social Networks*, pages 1–6, 2020.

[4] S. Alshamrani, A. Abusnaina, M. Abuhamad, D. Nyang, and D. Mohaisen. Hate, obscenity, and insults: Measuring the exposure of children to inappropriate comments in youtube. *arXiv preprint arXiv:2103.09050*, 2021.

[5] S. Alshamrani, A. Abusnaina, and D. Mohaisen. Hiding in plain sight: A measurement and analysis of kids' exposure to malicious urls on youtube. In *The ACM/IEEE Workshop on Hot Topics on Web of Things*, pages 1–6, 2020.

[6] S. Bae, E. C. Sung, and O. Kwon. Accounting for social media effects to improve the accuracy of infection models: combatting the COVID-19 pandemic and infodemic. *Eur. J. Inf. Syst.*, 30(3):342–355, 2021.

[7] bbc. Coronavirus: Victoria declares state of disaster after spike in cases, 2020.

[8] M. Bhat, M. Qadri, M. K. Noor-ul Asrar Beg, N. Ahanger, and B. Agarwal. Sentiment analysis of social media response on the covid19 outbreak. *Brain, Behavior, and Immunity*, 87:136, 2020.

[9] D. M. Blei and J. D. Lafferty. Topic models. In *Text mining*, pages 101–124. 2009.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[11] J. Brailovskaia and J. Margraf. The relationship between burden caused by coronavirus (covid-19), addictive social media use, sense of control and anxiety. *Comput. Hum. Behav.*, 119:106720, 2021.

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

[13] CommonSenseMedia. Common sense media, 2020.

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

[15] Developers. Webshrinker. https://www.webshrinker.com/, 2020.

[16] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[17] N. Diakopoulos and M. Naaman. Towards quality discourse in online news comments. In *Proceedings of the 2011 ACM Conference on Computer Supported Cooperative Work, CSCW 2011, Hangzhou, China, March 19-23, 2011*, pages 133–142, 2011.

[18] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 29–30, 2015.

[19] D. Domalewska. An analysis of COVID-19 economic measures and attitudes: evidence from social media mining. *J. Big Data*, 8(1):1–14, 2021.

[20] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

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

[22] P. Eachempati, P. R. Srivastava, and Z. J. Zhang. Gauging opinions about the COVID-19: a multi-channel social media approach. *Enterp. Inf. Syst.*, 15(6):794–828, 2021.

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

[24] FamilyZone. Familyzone, 2020.

[25] F. Figueiredo, F. Benevenuto, and J. M. Almeida. The tube over time: characterizing popularity growth of youtube videos. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China*, pages 745–754, 2011.

[26] U. Gasser, S. Cortesi, M. M. Malik, and A. Lee. Youth and digital media: From credibility to information quality. *Berkman Center Research Publication*, (2012-1), 2012.

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

[28] Haccer. TWINT - Twitter Intelligence Tool, 2020.

[29] IMDB. Imdb, 2020.

[30] J. Clement. Leading countries based on number of twitter users as of july 2020, 2020.

[31] Jeff Cox. IUS payrolls plunge 701,000 in March amid the start of a job market collapse, 2020.

[32] R. Kaushal, S. Saha, P. Bajaj, and P. Kumaraguru. Kidstube: Detection, characterization and analysis of child unsafe content & promoters on youtube. In *the 14th Annual Conference on Privacy, Security and Trust (PST)*, pages 157–164, 2016.

[33] KazAnova. Sentiment140 dataset with 1.6 million tweets, 2020.

[34] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour. Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter. *Cureus*, 12(3), 2020.

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

[36] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284, 1998.

[37] Y. Li, J. Shin, J. Sun, H. M. Kim, Y. Qu, and A. Yang. Organizational sensemaking in tough times: The ecology of ngos' COVID-19 issue discourse communities on social media. *Comput. Hum. Behav.*, 122:106838, 2021.

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

[39] Z. Ma, A. Sun, Q. Yuan, and G. Cong. Topic-driven reader comments summarization. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 265–274, 2012.

[40] Mark Mather. Life on Hold: How the Coronavirus Is Affecting Young People's Major Life Decisions, 2021.

[41] T. M. Massaro. Equality and freedom of expression: The hate speech dilemma. 1990.

[42] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[43] Mike Snider. Two-thirds of Americans, 227 million, play video games. For many games were an escape, stress relief in pandemic, 2020.

[44] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[45] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[46] MushroomNetworks. How sd-wan/multi-wan technology handles the data avalanche from youtube (infographic), 2020.

[47] L. Nemes and A. Kiss. Social media sentiment analysis based on COVID-19. *J. Inf. Telecommun.*, 5(1):1–15, 2021.

[48] A. Obadimu, E. Mead, M. N. Hussain, and N. Agarwal. Identifying toxicity within youtube video comment. In R. Thomson, H. Bisgin, C. L. Dancy, and A. Hyder, editors, *Social, Cultural, and Behavioral Modeling - 12th International Conference, SBP-BRiMS 2019, Washington, DC, USA, July 9-12, 2019, Proceedings*, volume 11549 of *Lecture Notes in Computer Science*, pages 214–223. Springer, 2019.

[49] G. S. O'Keeffe, K. Clarke-Pearson, et al. The impact of social media on children, adolescents, and families. *Pediatrics*, 127(4):800–804, 2011.

[50] A. Olteanu, K. Talamadupula, and K. R. Varshney. The limits of abstract evaluation metrics: The case of hate speech detection. In *Proceedings of the 2017 ACM on Web Science Conference, WebSci 2017, Troy, NY, USA, June 25 - 28, 2017*, pages 405–406, 2017.

[51] Ongig. A list of minority groups. https://blog.ongig.com/diversity-and-inclusion/minority-groups-in-america/, 2021. Accessed: 2021-10-01.

[52] X. Pan, D. M. Ojcius, T. Gao, Z. Li, and C. Pan. Lessons learned from the 2019-ncov epidemic on prevention of future infectious diseases. *Microbes and infection*, 22(2):86–91, 2020.

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

[54] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543, 2014.

[55] Pete Evans. Canada lost more than 1 million jobs last month as COVID-19 struck, 2020.

[56] A. R. Rahmanti, D. N. A. Ningrum, L. Lazuardi, H. Yang, and Y. J. Li. Social media data analytics for outbreak risk communication: Public attention on the "new normal" during the COVID-19 pandemic in indonesia. *Comput. Methods Programs Biomed.*, 205:106083, 2021.

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

[58] Ranker. Ranker, 2020.

[59] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

[60] M. Rights. Minorities and indigenous peoples. https://minorityrights.org/country/united-states-of-america/, 2021. Accessed: 2021-10-01.

[61] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*, 2017.

[62] Ruth Thorlby, Caroline Fraser, Tim Gardner. Non-COVID-19 NHS care during the pandemic, 2020.

[63] M. Schuster and K. Nakajima. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*, pages 5149–5152. IEEE, 2012.

[64] S. Shtovba, O. Shtovba, and M. Petrychko. Detection of social network toxic comments with usage of syntactic dependencies in the sentences. In *Proceedings of the Second International Workshop on Computer Modeling and Intelligent Systems (CMIS-2019), Zaporizhzhia, Ukraine, April 15-19, 2019*, pages 313–323, 2019.

[65] D. K. Sil, S. H. Sengamedu, and C. Bhattacharyya. Supervised matching of comments with news article segments. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 2125–2128, 2011.

[66] L. A. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber. Analyzing the targets of hate in online social media. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016*, pages 687–690, 2016.

[67] A. Smith and M. Anderson. Social media use in 2018. `tinyurl.com/y3htxhlq`, 2018. Accessed: 2019-09-09.

[68] A. Smith, S. Toor, and P. V. Kessel. Many turn to youtube for children's content, news, how-to lessons. `tinyurl.com/yx9526fx`, 2018. Accessed: 2019-09-09.

[69] snowballstem. `https://snowballstem.org/`, 2019. Accessed: 2019-05-01.

[70] S. O. Sood, J. Antin, and E. F. Churchill. Using crowdsourcing to improve profanity detection. In *Wisdom of the Crowd, Papers from the 2012 AAAI Spring Symposium, Palo Alto, California, USA, March 26-28, 2012*, 2012.

[71] Statista Research Department. Social media use during COVID-19 worldwide - statistics facts, 2021.

[72] J. Sun and P. A. Gloor. Assessing the predictive power of online social media to analyze COVID-19 outbreaks in the 50 U.S. states. *Future Internet*, 13(7):184, 2021.

[73] R. Tahir, F. Ahmed, H. Saeed, S. Ali, F. Zaffar, and C. Wilson. Bringing the kid back into youtube kids: Detecting inappropriate content on video streaming platforms. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019.

[74] Z. Tang, A. Miller, Z. Zhou, and M. Warkentin. Does government social media promote users' information security behavior towards COVID-19 scams? cultivation effects and protective motivations. *Gov. Inf. Q.*, 38(2):101572, 2021.

[75] Tara John and Ben Wedeman. Italy prohibits travel and cancels all public events in its northern region to contain coronavirus, 2021.

[76] The Atlantic. The public deserves the most complete data available about covid-19 in the us. no official source is providing it, so we are., 2020.

[77] A. Tommasel, J. A. Diaz-Pace, J. M. Rodriguez, and D. Godoy. Capturing social media expressions during the COVID-19 pandemic in argentina and forecasting mental health and emotions. *CoRR*, abs/2101.04540, 2021.

[78] M. Tsagkias, W. Weerkamp, and M. de Rijke. Predicting the volume of comments on online news stories. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 1765–1768, 2009.

[79] A. Vaidya, F. Mai, and Y. Ning. Empirical analysis of multi-task learning for reducing model bias in toxic comment detection. *CoRR*, abs/1909.09758, 2019.

[80] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *NeuRIPS*, 2017.

[81] Viji. Overview of minority communities in india. https://vikaspedia.in/social-welfare/minority-welfare-1/overview-of-minority-communities-in-india, 2021. Accessed: 2021-10-01.

[82] VirusTotal. VirusTotal, 2020.

[83] wikipedia. https://en.wikipedia.org/wiki/List_of_world_news_channels, 2019. Accessed: 2019-09-09.

[84] World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020, 2020.

[85] WorldOMeter. COVID-19 Coronavirus Pandemic, 2020.

[86] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[87] YouTube. https://developers.google.com/youtube/v3/getting-started, 2019. Accessed: 2019-05-01.

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

[89] M. Ziegele, T. Breiner, and O. Quiring. What creates interactivity in online news discussions? an exploratory analysis of discussion factors in user comments on news items. *Journal of Communication*, 64(6):1111–1138, 2014.