Building A Corpus For Toxic Texts in Live Stream Chat Logs

Phoebe J. Chen
phoebe_chen@dlsu.edu.ph
De La Salle University-Manila

Kyle Isaac S. Tagle
kyle_tagle@dlsu.edu.ph
De La Salle University-Manila

Johannes Aaron T. Uy
johannes_uy@dlsu.edu.ph
De La Salle University-Manila

Jerick John D. Wong
jerick_wong@dlsu.edu.ph
De La Salle University-Manila

ABSTRACT

Moderating toxicity in streaming platforms like YouTube can be challenging. In recent years, there are only a few studies that have analyzed toxicity in streaming platforms. Additionally, because toxicity can manifest differently on different social networking platforms, the definition for toxicity could differ and be unclear, which makes modeling toxicity for automated detection and analysis even more difficult. In this research, we will first define toxic communcation. We will then create a dataset and attempt to analyze toxicity in YouTube's livestream chat logs. In the next step, we will collect chat logs from targeted YouTube livestreams and annotate the data. After this, we will process the chat logs and extract textual features of the toxic messages. These features may potentially help with creating a a machine learning model which can then be used to detect toxicity that is present in livestream chat logs. This can possibly reduce the prevalence of toxicity in streaming platforms.

1 INTRODUCTION

In 2020, with the COVID-19 pandemic causing community lockdowns, statistics show that people have been tuning into online streaming platforms and services more. Statista reports that YouTube Gaming Live has averaged nearly 196,000 concurrent viewers in the second quarter of 2018 and nearly 759,000 concurrent viewers in the third quarter of 2020[1]. According to TwitchTracker, the hours watched on Twitch spiked from 880 million hours in both November and December 2019 to nearly 1.8 billion hours in April 2020 and 1.875 billion hours in December 2020[2]. [14] of BBC News noted that the average viewing time of a person in streaming services during the pandemic was about one hour and 11 minutes, including 12 million new users subscribing to the services. Additionally, [21] also noted that live streaming platforms Twitch, Facebook Gaming, and YouTube Gaming experienced a very significant growth in viewership, with Twitch gaining a total of 334 million hours of viewing time, Facebook with a 72 percent growth rate, and YouTube with 14 percent growth rate in between March and April.

One of the reasons why people watch streams is because of the direct connection it provides to its audience [20]. Live streams have a chat feature where the live audience can interact with the streamers or the streamed video. Although the chat feature allows direct and interactive experience for the audience and streamers, it may open to toxic behavior. People may show hostile behavior such as flaming, aggression, verbal abuse, and use of offensive language, which are disruptive [16, 17]. Audience in YouTube news channels that discuss sensitive topics such as racism, politics, and religion tends to show toxic behaviors such as cyberbullying, creating online firestorms, and expressing hate speech [18]. Because of this, managing or moderating chats in livestreams is important.

While people believe that enforcing chat rules by moderators and educating the audience of the rules are effective strategies to manage toxicity [4], large amounts of data can still be hard to process by humans alone. Moreover, human moderators can also be susceptible to emotional distress when repeatedly exposed to negativity [23]. Automated solutions can provide a better alternative for moderating chats.

Several studies have modeled online toxicity for automated detection [3, 7, 11]. Microblogging platforms like Twitter have many studies that modeled toxic content such as harassment, cyberbullying and hate speech. However, these models may not be applicable when processing livestream chats because the data in these types of the platforms are different. Data in live stream chats are much shorter and contains much more repetitiveness [22], as seen in figure 1.

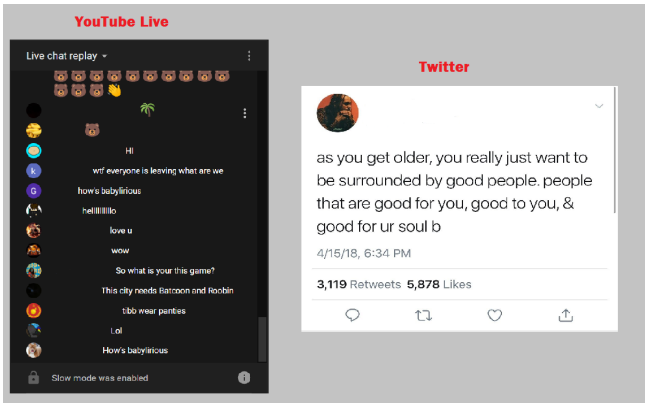

Figure 1: Chat Logs vs. Tweets

[1]https://www.statista.com/statistics/761100/average-number-streamers-on-youtube-gaming-live-and-twitch/
[2]https://twitchtracker.com/statistics

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Unfortunately, there are only a few studies that have processed data in streaming platforms. To the best of our knowledge and research, the chat feature in YouTube Live has none. This provides our study an opportunity to explore toxicity using the data of YouTube Live chat.

2 BACKGROUND AND RELATED WORK

In one of the studies done by Bosco et al., they discussed corpora development that can be used for sentiment analysis. For their data collection, they gathered texts from web services that provide product reviews such as Amazon. They also collected texts from micro-blogging websites such Facebook and Twitter. During their data collection, the following were kept in mind: the metadata of the posts, the genre of the post, and the topic. For their annotation process, they followed the guidelines of Cohen's kappa coefficient and used the annotation model provided by Turin University Treebank which includes tokenization and morphological and syntactic analysis. For the analysis of their corpora, the suggested method to check the annotation of the corpus is to compare human annotation versus automated machine annotation.

Similarly, Poletto et al. conducted a study that focused on building a corpus for Hate Speech in Twitter. In their research, they conducted it in phases which are data collection and data annotation. For their data collection, they first collected Italian Twitter posts. They then filtered the words into three categories: ethnic group, religion, and Roma. For their annotation process, they had to formally define hate speech. This was done in order for the annotators have a clear definition of hate speech. They also categorized other tweets as aggressive, offensive, and stereotype. They had five annotators in total. At the end of their research, they were only able to annotate around 1,500 tweets out of the 300,000 tweets that were collected.

In another related work that relates to corpus development is the work done by Filatova created a corpus using Amazon Reviews. Their goal was to create a corpus that identifies sarcastic or ironic reviews. They collected data through the platform Amazon Mechanical Turk. After their data collection, they needed to clean the process they received. They performed some text processing techniques in order to do so. For their annotation phase, they got annotators and decided to label though majority vote or using an data quality algorithm that was based on Krippendorff's alpha coefficient. Using that algorithm, they were able to determine whether the annotated labels were reliable.

3 METHODOLOGY/DESIGN

3.1 Data Collection

As mentioned in previous sections, a dataset for live stream chat logs are scarce. To resolve this issue, we decided to collect our own chat logs from YouTube. We collected from two primary genres on YouTube, gaming channels and news reporting channels. We only collected from the top channels of each genre. In order to determine the top channels, we utilized existing tools that are able to track the statistics and analytics of content creators on YouTube. One of these tools is the website called SocialBlade. SocialBlade is able to get these statistics and analytics on content creators on YouTube.

For us to find the channels that fit our criteria, we utilized the website's manual search feature. The website allows us to click the top lists per platforms which organizes the list by SocialBlade rank. The SocialBlade rank is a scoring metric that the website provides that into account how much views a channel is getting and how influential the channel is getting. This, however, will not fit in our criteria, as we do not know whether the SocialBlade ranking score is accurate, so in our search criteria we will be sorting the list by subscriber count. The website also gives us features that lets us query content creator per tag. These tags could include which country they are located in, what type of content we can find in their channel such as vehicles, comedy, education, entertainment, or film.

After finding channels that fit our criteria, we selected past live streams to collect our chat logs from. In order to collect the chat logs, we used an existing chat log collecting tool. The tool that we used is called Chat Replay Downloader. This tool was built on python and is an open source tool that we found in GitHub[3] that is able to collect chat logs from YouTube. This tool only requires the live stream URL and it will output a file of chat logs. The file will have the following columns:

Table 1: File Table Details

Column Name	Data Type
time_in_seconds	int
message	string
author	string
time_text	string
timestamp	string

3.2 Data Cleaning

After data collection, we were left with multiple CSV files. This means that the data we have collected are still unorganized. To organize all the collected data, we grouped together each file by channel inside folders. After organizing the chat logs together, we anonymized the users found in the chat logs. In order to anonymize the user we implemented our own anonymization algorithm. The output of our algorithm will format the authors name as "user@characters".

After the anonymization step, we removed the unneeded columns from the output of the chat collection tool. Then we added our own labels to the files. The final file for annotation will have the following columns:

Table 2: Data Table for Annotation Details

Column Name	Data Type
message	string
author	string
time_text	string
Direct Harassment	int
Hate Speech	int
Sexual Harassment	int
Trolling	int
Others	int
Toxic	int

[3]https://github.com/xenova/chat-replay-downloader

3.3 Annotation Labels

From various sources, toxic communication can be defined in multiple ways. One can define it as Flaming, which is when a person shows aggressive behavior towards a certain individual on the internet [10]. Using offensive language, trolling, or verbal abuse may also fall under the definition of toxic communication [12]. Because of these mixed definitions, we decided to add labels of Gabriel et al. to our corpus to determine whether these labels fall under the definition of Toxic or if they are correlated to each other.

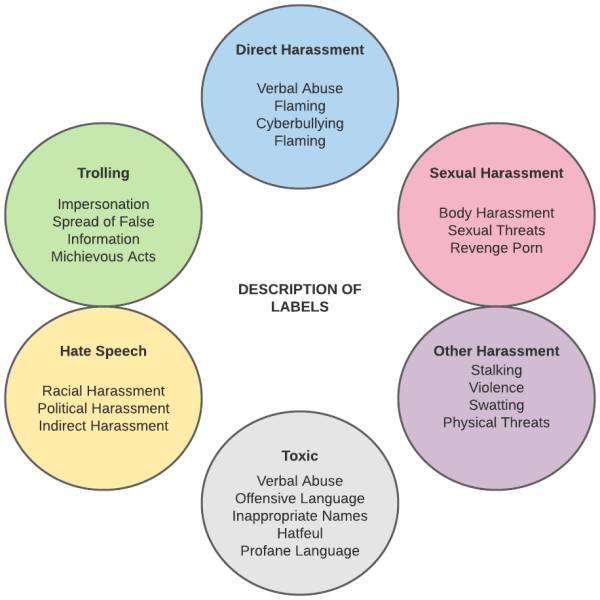

Figure 2

In Figure 2, we summarized all the possible definitions of each label. These definitions were created in combination with the definitions of the journal article of Gabriel et al.. In total, we have six labels which are *Direct Harassment*, *Hate Speech*, *Sexual Harassment*, *Trolling*, *Trolling*, *Others*, and *Toxic*. The first five labels are defined as follows:

Direct Harassment by definition is the act of speaking to a person or a group with the intention of bringing them down or upsetting them. Because it directly affects the mental health of an individual, this type of harassment can bring stress, anxiety, or depression [8]. There are different ways Direct Harassment can be manifested in text. One of these are Verbal Abuse, Flaming, and Cyberbullying.

Hate Speech is speech that shows hostility or is meant to be negative, humiliating, or insulting to members of a targeted group. *Racial* and *Political Harassment* belongs to the Hate Speech category. Hateful words or speech targets a group of people by their color, country, culture, and faith is considered racial harassment [5, 15]. An example of this would be the use of the word "n*gga". Political harassment targets people's political opinions. Political opinions could involve global warming, gun control, or the opioid epidemic [15].

Sexual Harassment occurs when a person online receives unsolicited sexual content [8]. A type of sexual harassment is *Body Harassment*. Body harassment occurs when a person treats a woman as a sex object. This type of harassment would include tags that would describe the woman's weight, breast size, and attractiveness or lack of attractiveness. For example, "-I cannot stop looking at Nikki's dreadful black crooked bra" [19].

Trolling is the act of eliciting pointless arguments, or using information incorrectly to encourage others to participate in useless, meaningless, and time-consuming discussions, even leading to negative violent or behavioral responses. We should view trolling as "deliberate, deceptive, and mischievous" acts with a purpose to trigger a response from the victim/s[9]. Cheng et al. mentioned that they considered griefing, swearing, or personal attacks as trolling. The act of flaming a person online can be also be considered as Trolling [8].

Others are texts that was found not to belong to any of the previous category will be placed here. An example of texts that do not belong in the previous categories would be *Stalking, Violence, Swatting,* and *Physical Threats.*

3.4 Annotation Process

For the creation of our labeled corpus, we hired three annotators for each batch of annotations. The annotators we hired had the following criteria:

- at least the age of eighteen (18) years old
- has access to a spreadsheet tool (Microsoft Excel or Google Sheets)
- must have access to an internet connection

3.4.1 Annotation Recruitment. For the recruitment of annotators, we followed a convenient sampling method. A poster was made and posted on Facebook and on the DLSU Community Forum Facebook group. We included alongside the poster a link to our Google Forms containing the sign up sheet for people who were interested in participating the annotation process of our research. A total of 38 people signed up for the annotation process, but only 17 of them ended up participating. Consent forms and annotation guidelines were sent to the 17 annotators. After signing the consent form, a folder containing the chat logs is shared to the annotators. As stated in the consent form, the annotators are to complete the annotation within one week after the chat logs are shared. After completing the annotations, the annotators are given an option to annotate another batch of chat logs. If they are still interested, they will have to sign another consent form before proceeding with the annotations.

3.5 Data Validation

For the final corpus, we created a design scheme to determine the final labels for each chat log. Because we had three annotators for each chat log, we decided that we will only keep the labels that were agreed upon by two annotators. We decided to keep it at

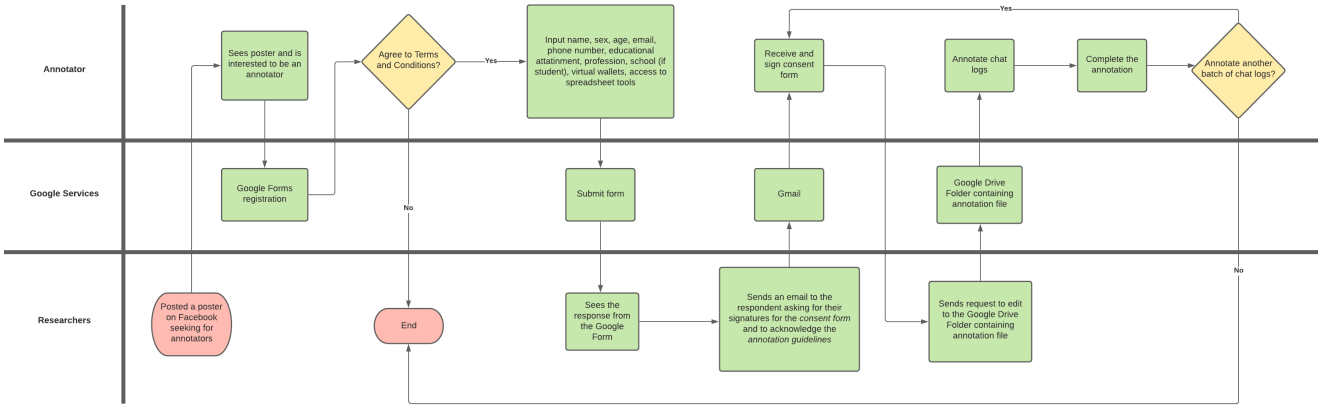

Figure 3: Annotation Process Flow

two because we determined that two is the majority vote for three annotators.

We also wanted to compute for the level of agreement of the three annotators. To do this, we used **Fleiss Kappa**. The formula for this is as follows:

$$k = \frac{P_a - P_s}{1 - P_s} \tag{1}$$

This formula will output a value that ranges from 0.00 to 1.00. These ranges can be interpreted in six ways as seen in Table 3 [8].

Table 3: Interpretation Table for Kappa Values

Kappa	Interpretation
k <0	Poor Agreement
0.01 - 0.20	Slight Agreement
0.21 - 0.40	Fair Agreement
0.41 - 0.60	Moderate Agreement
0.61 - 0.80	Substantial Agreement
0.81 - 1.00	Almost Perfect Agreement

3.6 Corpus Analysis and Evaluation

Term Frequency - Inverse Document Frequency A technique for data analysis and feature extraction is Term Frequency-Inverse Document Frequency (TF-IDF). This technique is a statistical approach to putting weight into a word's relevance to document [13]. It is also one of the most common approaches to putting weight in words in relation to their documents [1]. Using this technique, we were able to determine words that may appear toxic. The computational formula for TF-IDF is as follows:

$$TF - IDF = TF * IDF \tag{2}$$

where

$$TF = \frac{w}{n} \tag{3}$$

such that w for the number of times the word occurred in the document, and n for the total number of texts in the document.

and

$$IDF = log_E \frac{D}{W} \tag{4}$$

such that D is the total number of documents and W is the number of times the word has occurred at least once in of those documents. For ease of computation, we used TFIDFVectorizer found in scikit-learn, a python library, to get the TF-IDF scores of each n-gram.

Correlation Analysis Using the TF-IDF scores and the labels of the corpus, we ran a Correlation Analysis script to determine whether words or labels are correlated to each other. In order to do Correlation Analysis, we had to first compute for the Correlation Coefficients which has the following formula:

$$\rho = 1 - \frac{6^* \Sigma d_i^2}{n(n^2 - 1)} \tag{5}$$

The formula will output a number that ranges from -1 to +1, where -1 implies that there is a negative correlation between the two variables and +1 implies that there is a perfect relationship between the two variables. If the analysis results in a perfect zero, this implies there is no correlation between the two variables at all [2]. This can be seen in Table 4.

Table 4: Correlation Values Interpretation

Coefficient Value (+/-)	Interpretation
1.00	Perfect Correlation
0.70-0.99	Strong Correlation
0.40-0.69	Moderate Correlation
0.10-0.39	Weak Correlation
0.00	No Correlation

4 RESULTS

4.1 Annotators Recruited

In this section, we discuss the demographics of the annotators who participated in this study in order to be clear of possible biases. As seen in Figure 4, The annotators had a very even distribution with eight Male annotators and nine Female annotators.

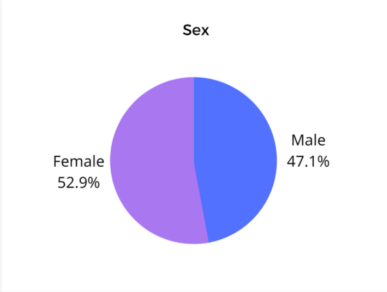

Figure 4: Number of Annotators based on Sex

Figure 6 shows a distribution of the annotators by age. For ages 20 and 21, there were five annotators for each. The remaining seven annotators were split among ages 18, 19, 22, and 27.

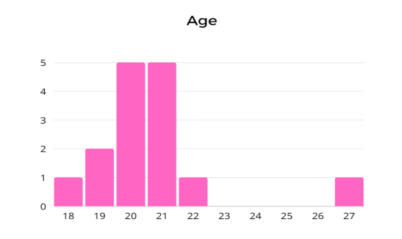

Figure 5: Number of Annotators based on Age

Figure 6 shows a distribution of the annotators based on the university they attended. It can be seen that majority of the annotators are students from De La Salle University (DLSU). This is most likely caused by the recruitment poster being posted on the Facebook group *DLSU Community Forum*.

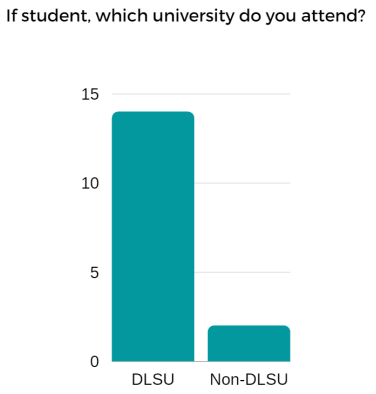

Figure 6: Number of Annotators based on University Attended

4.2 Annotated Data

Our data annotation process went on for a total of 4 weeks, which resulted into a total of 14 batches of chat logs annotated, where each batch contains a maximum of 3000 chat logs. The total amount of chat logs annotated is 37405.

We generated kappa scores based on each of the labels annotated per batch using Natural Language Toolkit's (NLTK) multi_annotator function. According to NLTK's documentation, this function uses the Fleiss Kappa algorithm to compute for the scores. Table X below shows the corresponding kappa score results.

Table 5: Annotated Data Info

Batch	Source	Chat Logs	DH	HS	SH	Tr	O	Tx
001	USA Today	3000	255	435	20	734	50	254
002	USA Today	3000	390	509	47	396	41	685
003	USA Today	3000	133	277	19	742	54	202
004	USA Today	3000	396	292	24	403	73	265
005	USA Today	3000	396	360	6	286	12	433
006A	Fox News	379	73	42	0	50	6	14
006B	Fox News	65	13	6	1	13	0	1
006C	Fox News	467	63	22	1	30	4	17
006D	Fox News	83	16	7	0	14	0	4
006E	USA Today	1683	63	89	1	56	2	20
007	PewDiePie	3000	29	8	41	14	6	44
008	Nogla	2000	77	19	7	24	0	22
009	PewDiePie	3000	14	6	7	15	1	21
010	Nogla	2000	88	77	14	54	0	73
011	USA Today	3000	166	178	5	191	24	46
012	PewDiePie	1898	82	20	2	97	3	8
013	Nogla	2931	161	77	8	68	13	53
014	PewDiePie	1899	43	4	6	43	8	15

4.2.1 Kappa Scores. After getting all the annotations done, we wanted to get the kappa scores of annotations. Using the kappa scores, we will get the agreement scores between the three annotators that was tasked to annotate the batch. These scores are important because these will initially tell us how much of the annotations will remain after getting majority vote. In Table 6, we calculated the kappa scores for each batch of annotation and the averaged their results per channel. As seen in the table, because of the low scores, we were able to assume that majority of the annotations will be discarded because of the low scores.

Table 6: Partial Analysis Results - Average of Kappa Results

Video	DH	HS	SH	Tr	O	Tx
PewDiePie	0.0197	0.0009	0.0277	0.2730	0.0000	0.0510
Nogla	0.0449	0.4167	0.1000	0.07484	0.6667	0.1090
USA Today	0.1034	0.2628	0.1891	0.2853	0.1465	0.2332

We also wanted to compute the scores the kappa scores per video, as seen in Table 7. To see if there is a difference in scores when done differently. But as seen in the table, it can still be concluded that the annotators only slightly agreed in their annotations.

Table 7: Partial Analysis Results - Kappa Results per Video

Video	DH	HS	SH	Tr	O	Tx
PewDiePie	0.03185	0.0001	0.0064	0.0607	0.0004	0.0761
Nogla	0.0331	0.5197	0.1207	0.1091	0.0000	0.1125
USA Today	0.3187	0.2841	0.2395	0.2831	0.2235	0.1799

4.2.2 Majority Votes. After getting the kappa scores, we ran a script that would get the majority vote of each dataset. From PewDiePie's video, only 29 chat log messages remained after getting the majority vote. For Nogla and USA Today's videos, 171 and 1422 chat logs are retained respectively.

4.2.3 USA Today Analysis. The total number of chat messages annotated in the USA Today video is 16683. As seen in Figure 7, more than 15000 chat messages were not tagged as any of the labels. The highest count among the labels is Hate Speech, while the Toxic, Trolling, and Direct Harassment labels had similar numbers. The label with the lowest count is Sexual Harassment.

	count
none	15241.0
Hate Speech	739.0
Toxic	399.0
Trolling	380.0
Direct Harassment	364.0
Others	65.0
Sexual Harassment	24.0

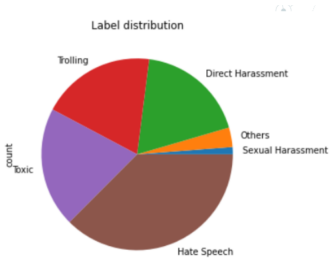

Figure 7: USA Today Annotated Chats

4.3 Analysis of Toxic Communication

After the collection of the annotated data, we performed several techniques to analyze and determine what toxic communication may look like in the live streaming space.

4.3.1 On News Genre. We performed a correlation analysis on the labels which can be seen on Figure 8. This is so that we can see if a message tagged in one label is likely to be tagged as other labels. There is fairly low correlation on the labels Which could mean that there is still a distinction between the labels. The label with the highest correlation to the Toxic label is Direct Harassment and Hate Speech, while sexual harassment messages were less likely to be tagged as toxic which may be due to the fact there is a low number of count in messages tagged as sexual harassment.

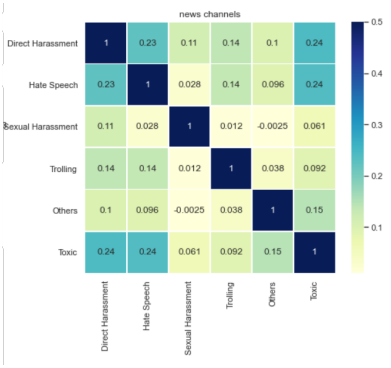

Figure 8: Word Cloud of Top Words for News Genre

From the results in Figure 9, we also wanted to determine what are the top words they may appear in each of the labels. We first extracted the top bi-grams from the news genre. From those results, we were able to observe that the top bi-grams had a lot of proper nouns. This indicates that when it comes to toxic communication in the news genre it tends to involve people of political power. Because the prominent amount of proper nouns, we re-ran our script, but consider the proper nouns as stop words. We did this to determine which words may actually appear in toxic. The results can be found in the following word cloud. From what is seen in the word cloud, it can be observed that toxic communication in the news has a lot of words that may be insulting to a person. For example, it can observed that the words “suck", “moron", and “idiot" is seen a lot.

Figure 9: Correlation Matrix for News Genre

From our initial run of analysis, we observed that there was a high repetition of the target’s name, name calling or insult involved, and the target was explicitly stated. Therefore, the next step to improving our results is to describe the dataset better by extracting other possible features by applying pre-processing techniques such as Part-of-speech (POS) tagging, and removing uppercase letters and punctuation marks.

Table 8: POS P-Value

Top Toxic	P-Value
NOUN NOUN INTJ	0.0704
ADJ NOUN NUM	0.1248
INTJ PROPN VERB	0.1248
ADJ PROPN ADV	0.1248
EMOJI NOUN VERB	0.1248

With the results in Table 8, we were able to determine what the syntax of toxic communication looks like in the news genre. However, as seen in the P-value scores, they are high. Ideally, in order for a hypothesis to be considered significant, The P-value must have a score less than 0.05.

After getting the scores of the POS sequences, we proceeded to get the results of our Topic Modeling. First we proceeded to compute the coherence score for the News Genre as seen in Figure 10

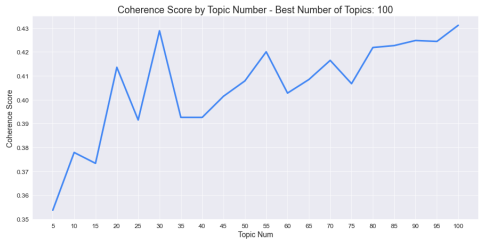

Figure 10: Coherence Score for News Genre

From Figure 10, we were able to determine the best number of topics to be extracted from the News Genre. In this case, it was 100 topics, as seen in Table 9

Table 9: Top Topics of Toxicity in News. Legend: *Tx* = Toxic; *DH* = Direct Harassment; *HS* = Hate Speech; *SH* = Sexual Harassment; *Tr* = Trolling; *O* = Others.

Topic	Topic Top 10 Keywords	Inferred Topic	DH	HS	SH	Tr	O	Tx
13	'win', 'must', 'winning', 'hope', 'big', 'crucial', 'wining', 'state', 'tired', 'wait'	the election	0.0266	0.049	-	0.0547	-	0.0252
14	'president', 'vice', 'mr', 'amazing', 'never', 'greatest', 'elect', 'bless', 'af1', 'next'	praising vice president	-0.0139	-	-	-	-	-0.0242
15	'america', 'first', 'bles', 'hate', 'wake', 'save', 'open', 'part', 'prosperous', 'burning'	mocking america	0.0273	0.0601	-	0.05	0.0244	0.023
19	'joe', 'sleepy', 'creepy', 'corrupt', 'george', 'dementia', '47', 'frack', 'hell', 'wake'	mocking Joe Biden	0.1278	0.1075	-	0.0953	0.0416	0.106
20	'thank', 'donation', 'kind', 'patriot', 'service', 'karen', 'side', 'mark', 'veteran', 'choose'	thanking patriots for service	-	-	-	0.0184	-	0.0173
24	'like', 'hit', 'button', 'smash', 'stream', 'share', 'look', 'depend', 'everyone', 'click'	show support to the stream	0.0628	0.07	0.0267	0.07	0.0351	0.0548
25	'california', 'dems', 'flip', '_rsbnredcalif', 'ruin', 'destructive', 'turn', 'latinos', 'yt', 'flipping'	latinos in california	-	0.0345	-	0.0377	0.0166	0.0182
29	'swamp', 'drain', 'integrity', 'bring', 'damn', 'corrupt', 'demonrats', 'evil', 'deep', 'back'	mocking democrats	0.0319	0.0353	-	0.0452	-	0.0159
33	'china', 'virus', 'pay', 'communist', 'home', 'must', 'owns', 'covid', 'electric', 'firing'	mocking china	0.0609	0.1355	-	0.1116	0.0491	0.0456
4	'pence', 'vice', 'indiana', 'speaker', 'yay', 'rock', 'think', 'prosperous', 'amazing', 'speech'	praising the speaker	-	-0.0246	-	-	-	-0.0174
41	'country', 'save', 'destroy', 'open', 'would', 'communist', 'across', 'god', 'leave', 'guns'	mocking the country	0.0541	0.0769	-	0.0667	0.0237	0.0331
42	'hunter', 'crack', 'whereâ', 'campaign', 'hunted', 'laptop', 'hell', 'footjobs', 'rose', 'barr'	mocking campaign	0.0566	0.0816	0.0214	0.0865	0.0414	0.0685
47	'suck', 'cnn', 'choice', 'song', 'news', 'by', 'fake', 'msm', 'twitter', 'netherlands'	mocking the news channel	0.069	0.0942	-	0.0843	0.0301	0.0809
48	'right', 'side', 'broadcasting', 'life', 'das', 'left', 'network', 'huh', 'uh', 'poll'	right side broadcasting	0.0361	0.061	-	0.0569	0.0283	0.0355
5	'biden', 'jail', 'rise', 'taxi', '47', 'gitmo', 'corrupt', 'moron', 'crime', 'home'	mocking Joe Biden	0.1727	0.1805	-	0.1713	0.0546	0.1455
52	'every', 'democrat', 'remove', 'state', 'time', 'votered', 'corrupt', 'count', 'friend', 'party'	mocking democrats	0.0413	0.0838	-	0.0778	0.029	0.0306
54	'lol', 'cry', 'talk', 'troll', 'live', 'mask', 'supporter', 'liberal', 'wear', 'hate'	mocking liberals	0.0736	0.0462	-	0.0546	0.0258	0.0509
57	'canada', 'germany', 'europe', 'today', 'austrian', 'russian', 'france', 'big', 'hope', 'terror'	terrorism	0.0241	0.0382	-	0.0439	0.0248	0.0334

Interesting insights can be gained from the extracted topics in the news genre.For example, as seen in Topic 14 of Table 9, topics that are related to the praising of the vice president is negatively correlated with toxic, having a score of -0.0242. This result is expected because giving compliments or praise to another person should not be considered toxic, as defined in our Chapter 3 and Annotation Guidelines. However, topics that involve mocking a target or subject as seen in Topics 15, 19, 29, 33, and 42 are topics that are positively correlated to the Toxic label. These were also positively correlated to other labels such as Direct Harassment, Hate Speech, Trolling and Others.

4.3.2 On Gaming Genre. In comparison to the News Genre, we were not able to collect as much as labelled data as we wanted. As seen in Table 10, we had a total of 33,068 chats to be annotated. But, after running them through our annotators, they we there was only 283 chat messages that had labels on them or only about 0.9% of the total number of chats. Even with a low amount of numbers, we still proceeded with the analysis of the genre.

Table 10: Toxic Communication on Gaming Genre. Legend: *DH* = Direct Harassment; *HS* = Hate Speech; *SH* = Sexual Harassment; *O* = Others; *Tx* = Toxic; *Tr* = Trolling.

# of Chats	216,750	
# of Annotated	33,068	15.3%
# of Labelled	283	0.9%
DH	70	24.7%
HS	126	44.5%
SH	26	4.6%
O	5	1.8%
Tx	51	18.0%
Tr	43	15.2%

From Figure 11, similar to that of the News genre, there is also a weak correlation between the Toxic label and the other labels in the Gaming genre. The correlation coefficients are as follows: 0.18 for Direct Harassment, 0.076 for Hate Speech, 0.041 for Sexual Harassment, 0.021 for Trolling, and 0.065 for Others. These results, however, may be subject to further research because of the low amount of labels collected.

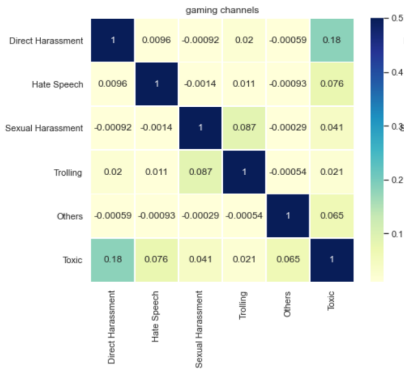

Figure 11: Correlation Matrix for Gaming Genre

In Figure 12, we were able to observe that the presence of profanity or vulgar words is more common compared to the news genre. In the news genre, the words used were negative, but they are not considered vulgar. For example, the words “fuck", “bitch", “fucker", and “stfu" are repeatedly mentioned in the toxic category.

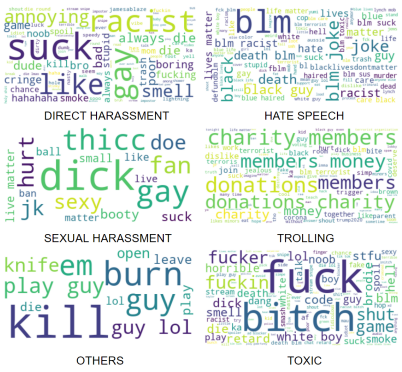

Figure 12: Correlation Matrix for Gaming Genre

From the results of the top words, we did not see a lot of proper nouns. However, we still proceeded with the extraction of POS Tags for each chat message. The following results can be seen in Table 11

Table 11: POS P-value

Top Toxic	P-Value
NOUN VERB PROPN	0.0704
VERB NOUN ADV	0.0704
NOUN ADV ADJ	0.0704
NUM NOIN	0.0704
PRON NOUN	0.1248

Similar to the results of the news genre, the P-value scores have a value that is less than 0.05. The values, however, are near the value. As discussed previously, the numbers of labelled chats in the gaming genre compared to the news genre is considerably lower.

With the increase of labelled data, it may be possible for the scores in Table 8 to decrease, and possibly be statistically significant.

When we proceeded to compute for the coherence score for the gaming genre, the results were not we expected. As seen in Figure 13, the highest coherence score is for five topics.

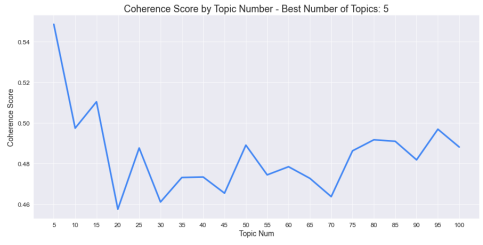

Figure 13: Coherence Score for Gaming Genre

Having the best number of topics to be modeled be five is not a good for us because the amount of topics would not have enough data for us to analyze. We then proceeded to check the next highest coherence score which was 15 topics. Similar to before, this amount of topics of data is not enough to perform analysis on. We then settled on getting 95 topics to be extracted as seen in Table 12

Table 12: Top Topics of Toxicity in Games.Legend: *Tx* = Toxic; *DH* = Direct Harassment; *HS* = Hate Speech; *SH* = Sexual Harassment; *Tr* = Trolling; *O* = Others.

Topic	Topic Top 10 Keywords	Inferred Topic	DH	HS	SH	Tr	O	Tx
11	'blm', 'joke', 'racist', 'death', 'talk', 'terrorist', 'stupid', 'stfu', 'toxic', 'acab']	racism	0.0269	0.1399	0.0155	0.0205	-	0.0369
17	'oh', 'uh', 'god', 'boy', 'yeah', 'shit', 'nooo', 'fuck', 'noooo', 'shoot']	negative interjections	-	-	-	-	-	0.0288
27	'ad', 'blocker', 'twitch', 'mobile', 'premium', 'sub', 'still', 'youtube', 'yt', 'nope']	blocking ads	-	-	-	-	-	0.0173
28	'stream', 'forget', 'enjoy', 'snipe', 'stop', 'start', 'among', 'live', 'look', 'politic']	stream snipe	0.0278	-	-	0.0257	-	0.0189
36	'gg', 'yay', 'win', 'lmfao', 'pete', 'lmaoooo', 'mod', 'team', 'ayyy', 'quick']	interjections	0.0117	-0.0163	-	-	-	0.0147
59	'matter', 'black', 'life', 'op', 'blue', 'live', 'ops', 'lives', 'white', 'racist']	black lives matter	0.0303	0.0262	0.0152	0.019	-	0.0252
75	'loud', 'af', 'asf', 'really', 'sean', 'little', 'always', 'wtf', 'enough', 'still']	negative interjections, loud stream	0.0147	-0.0108	-	-	-	0.0256

5 CONCLUSION AND FUTURE WORK

Online Toxicity has always been a problem in an online environment and there has already been several news articles about internet content creators and users of social media being victims of toxicity. While there already exists several studies about toxicity in an online environment, most of these studies were focused on long texts such as microblog posts. To the best of our research, there is no existing study of the presence toxic messages in live stream chat messages.

From the gathered related literature, one common insight that can be gained from it was that the definition of toxic messages varies. Because of this, we gathered all possible definitions of toxicity and incorporated labels (Direct Harassment, Sexual Harassment, Hate Speech, and Trolling) used by [8]. We proceeded to collect live stream chat logs from YouTube and were able to gather a total of 257,110 chat messages from different channels of the News and Gaming genres. In order to annotate these chat messages, we hired annotators and created an annotation guideline that would guide them in their work. We opened our annotation tasks to any person that was at least 18 years old.

From the annotations submitted by our annotators and getting the annotation agreement scores, we were able to create several multi-labeled corpora, with each corpora having a name linked to the channel from where we collected the live stream chat logs. We have corpora from the following YouTube channels: USA TODAY, Fox News, PewDiePie, Daithi de Nogla, Valkyrae, Terroriser, and DrDisRespect. The total chat messages that were collected and annotated from each YouTube channels are as follows: 19,683 from USA TODAY, 994 from Fox News, 9,797 from PewDiePie, 6,931 from Daithi de Nogla, 4,474 from Valkyrae, 5,866 from Terroriser, and 6,000 from DrDisrespect.

From the annotated corpora, we noticed that there were significantly more tags under the News genre compared to the Gaming genre. We could interpret that the News genre may be more toxic than the Gaming genre, however there also may have been some confusion between the annotators during the annotation of the gaming genre. Perhaps, there were some gaming terms that the annotators could not understand which led low amount of labels. This may be something to improve on in the annotation guidelines for future researchers.

After developing the corpora. we analyzed the labels that were assigned to each chat message. The first analysis we did was getting the TF-IDF scores of the n-grams in the chat logs. From the TF-IDF scores we were able to get the top words found in each label. We primarily focused on the top words under the toxic label and compared it whether the top words in the toxic label may also be found in other labels. We also computed for the correlation coefficients to determine whether there was a correlation between toxicity and the other labels, which was eventually found to be moderately correlated.

Aside from getting the top words through TF-IDF, we also performed POS tagging and extraction of POS sequences. We performed a correlational analysis with the POS sequences and the labels and were able to find that there is a correlation between the syntax of Toxic Communication and the other labels. From our results, most Toxic POS tags consistently had moderate positive correlation with Direct Harassment.

In addition to TF-IDF and Topic Modeling, using NMF was also a focus for this study. With topic modeling, inferred topic extracted from the collected corpora helped in identifying Toxic Communication in the two genres. To get the number of topics needed for each genre, the coherence score was calculated, and the best coherence score was used as basis to get the best number of topics that we needed.

After performing all these analysis techniques, it can be generalized that when it comes to Toxic Communication in the News genre, texts or words that are focused on putting a persons of political power or countries down are considered toxic. This includes mockery or derogatory terms. On the other hand, for the gaming genre, texts that vulgar or profanity in them are more likely to be considered toxic.

These analyses were based on the annotated data. Since convenience sampling was performed in order to recruit for these annotators. There may have been some bias incorporated in the corpora. Majority of our annotators were from DLSU and were in the age range of 18 years old to 25 years old. Their different backgrounds and experiences may have affected our corpora to some degree.

5.1 Future Work and Recommendations

In our research, we were able to annotate a few chat messages only out of the total chat messages that we collected. To further gain insights on Toxic Communication, we recommend future proponents that will continue our work to annotate more chat messages. In our research, we primarily focused on the collection of chat logs from the News and Gaming genre. We recommended future researchers to look into other genres such as Sports to compare whether Toxic Communication in other genres are similar or different. We also recommended to look into getting chat logs from channels of the same genre that have chat moderators and the others without, so that there is a comparison to identify how the presence of chat moderators can affect the number of toxic chat messages in each channel. From our results, we found that there was less toxic tags in the gaming genre compared to the news, which led us to assume that perhaps the Gaming genres were heavily moderated compared to the News genre.

Upon receiving the results of our study, we presumed that one possible reason for the Gaming channels to be less toxic than the News channels could be that the *streamer* or the owner of the channel is able to build a community of their own. Another reason could be that chat moderators might be more active in Gaming channels than in News channels. However in our research, we were unable to find basis for such claims. Therefore, we also recommend further investigation on why the Gaming genre is found to have less frequent Toxic Communication than the News genre.

Future researchers may also focus more on different types of analysis that can be performed on the chat messages. In this research, we only performed TF-IDF, POS Tagging, and Correlation Analysis. To further gain insight on these texts, perhaps more NLP techniques can be performed to better analyze what Toxic looks like in YouTube live streams. Future researchers may also focus more on cleaning the dataset. From our collected chat logs, we were able to observe that there tends to be a lot more spam messages in live stream chat logs. These spams may either be emojis or phrases.

The change of live stream platforms may also give more insight to future proponents. In our research, all the data were collected from YouTube. Future researchers may opt to collect data from other live streaming platforms like Twitch and Facebook Live. They can perform an analysis of toxic chats between the different platforms and find out if there are similarities or differences in toxic messages between the platforms.

To the future researchers that will continue our work, we recommend to improve upon our annotation guidelines. In our annotation guidelines, we did not provide the annotators much context to the video topics of the chat messages they were annotating. We recommended that the annotators must be given more context prior to annotating the chat messages. Requiring the annotators to watch the video prior to the annotation is a valid option. We also recommend that the assigned annotators must be experienced viewers of the live stream channel if they are to annotate the channel's live stream chat messages. This is because there may be terms in the chat that only experienced viewers may know full context of. Regarding the budget for hiring annotators, we recommend that the payment of the annotators that will be assigned to the Gaming genre be reconsidered for lesser compensation. From our research, we observed that toxic chats in the Gaming genre were much less frequent, which could make the annotation very much faster to accomplish than when annotating in the News genre where toxic chats were much more frequent.

REFERENCES

[1] Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39, 1 (2003), 45–65.

[2] Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish journal of emergency medicine* 18, 3 (2018), 91–93.

[3] C. Bosco, V. Patti, and A. Bolioli. 2013. Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT. *IEEE Intelligent Systems* 28, 2 (2013), 55–63. https://doi.org/10.1109/MIS.2013.28

[4] Jie Cai and Donghee Yvette Wohn. 2019. What Are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing (CSCW '19)*. Association for Computing Machinery, New York, NY, USA, 166–170. https://doi.org/10.1145/3311957.3359478

[5] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. 2012. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. 71–80. https://doi.org/10.1109/SocialCom-PASSAT.2012.55

[6] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1217–1230. https://doi.org/10.1145/2998181.2998213

[7] Elena Filatova. 2012. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing.. In *Lrec*. Citeseer, 392–398.

[8] Micah Gabriel, Plinky Gamara, Isabella Vicencio, and Jaymee Villacruz. 2020. *Building A Multi-label Online Harassment Corpus through Crowdsourcing*. Master's thesis.

[9] Maja Golf-Papez and Ekant Veer. 2017. Don't feed the trolling: rethinking how online trolling is being defined and combated. *Journal of Marketing Management* 33, 15-16 (2017), 1336–1354. https://doi.org/10.1080/0267257X.2017.1383298

[10] Noam Lapidot-Lefler and Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior* 28, 2 (2012), 434 – 443. https://doi.org/10.1016/j.chb.2011.10.014

[11] Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate speech annotation: Analysis of an italian twitter corpus. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, Vol. 2006. CEUR-WS, 1–6.

[12] Roman Poyane. 2020. Toxic Communication on Twitch.tv. Effect of a Streamer. In *Digital Transformation and Global Society*. Communications in Computer and Information Science, Vol. 1038. Springer International Publishing, Cham, 414–421.

[13] Shahzad Qaiser and Ramsha Ali. 2018. Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications* 181, 1 (2018), 25–29.

[14] Amol Rajan. 2020. TV watching and online streaming surge during lockdown. *BBC News* (Aug 2020). https://www.bbc.com/news/entertainment-arts-53637305

[15] Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L Shalin, and Amit Sheth. 2018. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*. 33–36.

[16] Rines. 2020. Online Community Bonds as a Method of Mitigating Toxicity in an Interactive Livestream. (2020). https://www.igi-global.com/chapter/online-community-bonds-as-a-method-of-mitigating-toxicity-in-an-interactive-livestream/253728

[17] T. Saarinen. 2017. Toxic Behavior in Online Games. (2017). http://jultika.oulu.fi/files/nbnfioulu-201706022379.pdf

[18] Joni Salminen, Sercan Sengün, Juan Corporan, Soon-gyo Jung, and Bernard J. Jansen. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLOS ONE* 15, 2 (02 2020), 1–24. https://doi.org/10.1371/journal.pone.0228723

[19] Sima Sharifirad and Stan Matwin. 2019. When a tweet is actually sexist. A more comprehensive classification of different online harassment categories and the challenges in NLP. (2019).

[20] M. Sjoblom and J. Hamari. 2016. Why do people watch others play video games? An empirical study on the motivations of Twitch users. (2016). https://www.sciencedirect.com/science/article/abs/pii/S0747563216307208

[21] Bijan Stephen. 2020. The lockdown live-streaming numbers are out, and they're huge. *The Verge* (May 2020). https://www.theverge.com/2020/5/13/21257227/coronavirus-streamelements-arsenalgg-twitch-youtube-livestream-numbers

[22] David C. Uthus and David W. Aha. 2013. Multiparticipant Chat Analysis: A Survey. *Artif. Intell.* 199–200, 1 (June 2013), 106–121. https://doi.org/10.1016/j.

artint.2013.02.004
[23] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300390

A ETHICAL CONSIDERATIONS

In order for our study to be valid and legal, we took note of several ethical issues that might arise, especially since our study is mainly focused on toxic messages.

When we collected chat logs from YouTube live streams, we made sure to anonymize the username of the author to protect the users. As the live stream chat logs may contain toxic, vulgar, and explicit words which may trigger emotional and mental distress, we forewarned the annotators through the consent form that we sent them that the chat logs that we will have them annotate may contain such features. In addition to forewarning, we only allowed annotators who are at least 18 years of age to participate in the annotation process. To ensure the security and confidentiality of the personal information submitted by the annotators, we will retain their data for no longer five years. The ethical issues involving the participation of the annotators were properly noted in the informed consent form that we sent them prior to their participation in our study.