2022 IEEE 22nd International Conference on Software Quality, Reliability, and Security Companion (QRS-C) | 979-8-3503-1991-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/QRS-C57518.2022.00012

Social Media Safety Practices and Flagging Sensitive Posts

Lisa M. DiSalvo[1], Gabriela Viviana Saenz[2], W. Eric Wong[2,*], and Dongcheng Li[2]
[1]Arcadia University, Glenside, Pennsylvania, USA
[2]University of Texas at Dallas, Richardson, Texas, USA
ldisalvo@arcadia.edu, gabriela.saenz@utdallas.edu, ewong@utdallas.edu, dxl170030@utdallas.edu
*corresponding author

Abstract—**Today, social media has uniquely become a force for positive change, community building, and sharing ideas. But with millions of people, of all ages, tuned into social media many hazards and mishaps can arise due to inadequate content monitoring. Therefore, our research problem deals with social media safety practices, flagging sensitive posts, and effective content monitoring. In this paper, an examination will be made of the most impactful hazards and mishaps that arise from poor social media content monitoring, incorrect flagging, public safety, and the spread of misinformation. In addition, this paper will discuss the shortcomings of content monitoring tools currently on the market and how they can be improved. Finally, this paper goes over the findings and areas of research related to the University of Texas Dallas 2022 Software Safety REU program. Through the synthesis and culmination of exploring social media safety practices, we have curated an application to manually label social media posts according to a pre-established corpus of violent phrases.**

Keywords-social media; safety; misinformation; content flagging; hazard

I. INTRODUCTION

Social media is a ubiquitous force for positive change and community building. However, with millions of people tuned into social media [1], many hazards and mishaps can arise due to inadequate content monitoring. Lack of effective safety protocols and flagging algorithms on social media platforms lead to incorrectly flagged or unflagged posts, the spread of misinformation, and threats to public safety.

Incorrect flagging occurs when a social media platform uses its built-in machine learning model to flag posts, typically under the categories of violence, hate speech, or bullying and harassment. If a post is flagged incorrectly, harmless content is removed from the site, resources are wasted on the effort and other harmful posts are allowed to remain on the site. Harmful posts on social media pose a threat to public safety, deriving from the post itself and the claims made by its writer or from the spread of misinformation. Violent perpetrators aim to incite fear, and when their posts go undetected the threat to public safety affects the real world. Positive communities on popular platforms help combat threats, but people in severe circumstances cannot get the help they need if they are clouded with misinformation. Misinformation is shared in mass online. In the cases of violent acts, local breaking news, or natural disasters, users must be able to discern accurately from inaccurate information. These mishaps are illustrated in world events such as the school shooting in Uvalde, TX [2], the ongoing threat of cyber-bullying among the youth [3], and the spread of misinformation during the COVID-19 pandemic [4].

With these grave consequences, social media safety and content monitoring need to be reformed. Therefore, this research will analyze the shortcomings of content monitoring tools currently on the market and how they can be improved. To work toward a solution, we utilized Twitter APIs such as Tweepy [5] and Twint [6] to curate various functions that sort through a list of post images and text to label manually. The libraries used collect usernames, follower count, post like count, image links, date of posts, and hashtags used. This data aids in performing logistic regressions on the time frames and gathering statistical data for analysis. Our solution will scrape data from multiple social media platforms at once, collect text and images from every post, and use a machine learning model to properly categorize violent and non-violent posts. The results demonstrate the ability of machine learning to be used to create efficient systems of content monitoring with cross-platform abilities that prioritizes user safety. Future research includes testing machine learning models with the collected data to find the most effective algorithm and deploying a system with a user-friendly interface that will help prevent violence and maintain public safety.

II. RELATED STUDIES

Content Monitoring on social media comes with several potential mishaps and hazards. The most important and impactful mishaps and hazards include incorrect flagging, public safety, and the spread of misinformation. Incorrect flagging occurs when a social media platform uses its building machine learning model to flag posts, typically under the categories of violence, hate speech, or bullying and harassment [7]. If a post is flagged incorrectly, harmless content is taken down from the site, resources are wasted on this effort and other harmful posts are allowed to remain on the site. When harmful posts stay up on social media, a threat is posed to public safety, deriving from the post itself and the claims made by the writer of the post or from the spread of misinformation. Social media is often a tool used by violent perpetrators to incite fear in potential victims or to warn against a forthcoming crime they plan to commit.

If these posts go undetected, public safety is put at severe risk in the real world, not just the cyber world. One way to combat these issues is through positive communities that are

2693-9371/22/$31.00 ©2022 IEEE
DOI 10.1109/QRS-C57518.2022.00012

built on popular platforms, such as Twitter, Instagram, and Facebook. However, those in severe circumstances cannot get the help they need if they are clouded with misinformation. Misinformation is shared in mass online, and in the cases of violent acts, local breaking news, or natural disasters, users need to be able to discern accurate from inaccurate information, and the safety practices from content monitoring need to be able to perform these tasks [8]. Social Sentinel is a rising social media, content monitoring tool from the self-named company that is currently on the market. Social sentinel sells its software specifically to schools, typically high schools and middle schools with young social media users and uses its software to monitor the student's social media [9]. The hope in monitoring their content is to protect them from unnoticed signs of poor mental health, self-harm, or harm unto others and then provide them with the needed counseling at school. However, most teens have an online presence across multiple platforms, and Social Sentinel only monitors one platform at a time [10]. So, while a school feels students are safe based on their Twitter posts, they could be missing a completely different side of them on their Instagram posts. If a school wanted to monitor multiple platforms, they would have to take on additional charges, put in more resources, and waste more time going through each platform one at a time. In order to effectively prevent social media software mishaps, there needs to be a way to gather and monitor large amounts of social media posts at once across various platforms.

Moreover, The technical studies and ideas in this research endeavor specifically focus on the meshing of concepts presented in the following analysis papers, 'Use of a bot and content flags to limit the spread of misinformation among social networks: a behavior and attitude survey' [11], 'Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter' [4], 'The effectiveness of flagging content belonging to prominent individuals: The case of Donald Trump on Twitter' [12], 'Violent Political Rhetoric on Twitter' [13], 'HateCheck: Functional Tests for Hate Speech Detection Models' [14], and 'Offensive Language Detection in Nepali Social Media' [15]. To begin, Lanius, Weber, and MacKenzie Jr. discuss that 'The COVID19 crisis, which led to much of social life migrating online, has contributed to an infodemic' [11]. An infodemic describes the influx of information, specifically news being spread on all social media platforms, almost constantly. While this phenomenon can be associated with social media before and after the height of the pandemic, in this instance, the fear and uncertainty surrounding the COVID-19 virus led to harmful misinformation being spread, purposefully saturating social networks with false information, posing a significant real-life risk for users consuming this information. Furthermore, Lanius et al. [11] directly reference a study done in the work 'Coronavirus Goes Viral', by Kouzy et al. [4] (2020), which states that 'almost 25 percent of COVID19-related tweets contained some misinformation. Along with this statistic, Lanius et al [11]. highlight that even before the pandemic, misinformation spreads much easier opposed to the spread of factual information. Lanius [11] attributes the spread of misinformation in their research to Twitter bot accounts. These bot accounts are curated solely with the purpose of spreading misinformation through spam posting. Another method of misinformation spreading is highlighted in the research done by Chipidza et al [12]. and Kim [13]. Their studies delve into the role violent political information plays in the spread of misinformation on Twitter. Chipidza [12] notes that 'misinformation or fake news became increasingly salient following the 2016 presidential election in the United States. The surge of misinformation during the pandemic is very closely alike to the misinformation spread during the 2016 election. The 2016 election can be seen as the inception of social media becoming grounds for political propaganda [16], and the state of social media websites during the COVID-19 pandemic showcases how fake news techniques and strategies translated and evolved to promote fear and ultimately harm users on all platforms.

Furthermore, having noted previous statistics and tactics to which misinformation has been spread, as we emerge out of the heart of the COVID-19 pandemic, it is pivotal to understand that the spread of information has notably advanced. Users who seek to spread misinformation or create cyber-panic utilize their newfound knowledge gained during the height of the pandemic in current online scenarios. The state of social media and the internet has sociologically evolved into something much bigger, something that has a real-world impact on the world's economy, and most importantly, a user's safety.

Through a culmination of understanding how the spread of misinformation occurs on our target social media platform: Twitter, through content flagging and labeling along with sentiment analysis, prohibiting the spread of misinformation on Twitter is placed at the forefront. Our technical program is a tool that seeks to take into consideration the preconceptions of whatever the user specifies it to. Therefore, harmful, or deceitful information of all sorts can be removed. Finally, it is important to note that our tool cannot directly flag posts, it can only scrape posts, as our capability as researchers only allow us so much control over social media platforms.

III. SOCIAL MEDIA TRIGGER LABELING AND SAFETY PRACTICES

What is social media safety and how can a user strive to protect their information from unwanted access? Along with this, how can the spread of misinformation be mitigated, and overall prohibited from affecting a user's safety overall? Social media, almost akin to its real-life counterpart is a virtual hub or community in which users can share media related to their lives, whether it be for personal purposes or for business purposes. With the recent shift in the way social

media and technology interact with its users sociologically, it is important to note that social media, which was once intimate, is now a prime agent for consumerist advertising. The rise of consumerism and online shopping has directly interacted with the way advertisements and information is relayed. Everything on current social media platforms is attempting to garner a user's attention. This attention is now quantified as clicks, likes, or views. The spread of misinformation in a shocking manner, which may include the direct promotion of violent or disturbing content is dubbed 'click-bait'. As the name states, the post is meant to bait users to click on it due to the curiosity that arises from utilizing a shocking headline or image to promote a post. As highlighted by the research completed by Majid and Kouser [17], the following security issues and solutions are a compilation of common user mistakes which can lead to a security breach.

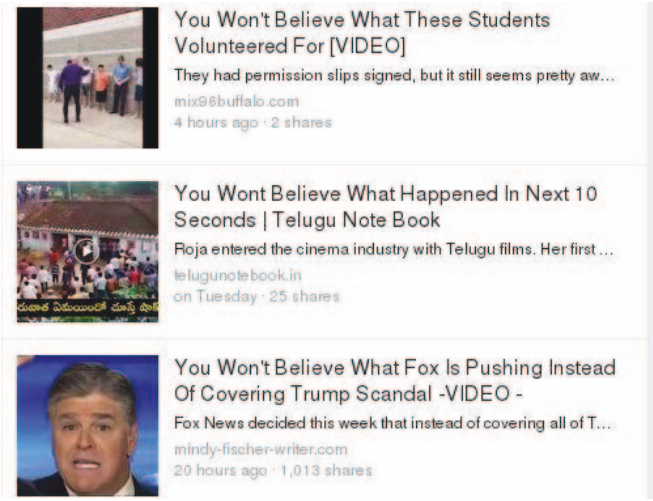

Figure 1. Example of Click-bait

A. Providing Login Information and Passwords to Fake Websites or Malicious Users

Many 'click-bait' websites utilize viral trend advertisements to market their online stores (as shown in Figure 1). On these online stores, you are enticed to purchase products, which then leads to you in putting your email, passwords, and even credit card information. In order to avoid this occurrence compromising a user's information, a user should ensure that the websites they are visiting are 'secured', which is usually highlighted by a lock symbol beside the hyperlink in a browser's search bar. Along with this precaution, users should use encrypted payment methods such as PayPal or cryptocurrency, therefore a user's actual payment method information is protected by the third-party application. Users should also not visit websites they are not familiar with, especially if said website utilizes 'eye-catching' terms and graphics. Malicious users on social media websites utilize bot accounts to spam message profiles with solicitations of sex or money. These malicious accounts bait regular users with links, which they will usually preface with an offer. For example, a malicious account could be sending you a link to "win a cash prize". Embedded in these links can be ransomware, viruses, or even tracking software. Users should completely avoid and block malicious unrecognized accounts and never click on a link they do not recognize. Overall, a user should not interact with advertisements or messages from companies or accounts they do not recognize.

B. Content Filtering on Social Media Applications

Many social media applications now provide users the ability to filter the type of content they want to see on their social media feed. Users are able to block accounts and restrict certain accounts and, in some instances, shocking posts are hidden automatically by the application to protect user safety. It is important that users set their content filtering specifications to their preferences. In addition, It is important to take into consideration if the source of content being promoted on a social media website is credible or not. Many social media platforms provide 'verification' status to celebrities, businesses, and large organizations to provide a virtual certificate of authenticity, essentially the verified check mark communicates to other users on a platform that 'this is the real deal. It is also important that users block and restrict accounts that spread misinformation, specifically if said accounts are not credible sources, businesses, or verified individuals.

C. Using Third-Party Apps

Third-party apps come as part of social media websites. These Third-Party Applications market themselves as 'follower tracking' or 'profile promotion' applications. Third-Party Applications are seen as essential for applications such as Instagram and TikTok, which are both platforms heavily focused on content engagement and quantitative statistics, which determines how 'good' your post is, or how many audiences it has reached. A user who seeks to use this kind of app needs to grant access to the third-party app, which provides said third-party application complete access to your account, even having access to sensitive information such as your password or credit card information. Not all third-party applications require you to log in on their platform with your social media account information. Many apps redirect users to their Instagram application to verify their account status in a more secure way. While third-party applications can be useful, the plethora of applications available on the App Store must mean that creators of these applications are profiting from not only the services provided to users but also from the mass intake of user information input into the application. Before downloading a Third-Party Application for a Social Media Website, users should perform thorough background research on the application they intend to use. Hence, if a user is to utilize a Third-Party Application, the user should refrain from entering their login information into the said application.

D. Passwords and Using Two Factor Authentication

Using the same password on multiple applications can lead to devastating results. If a user is compromised on one social media, account and utilizes the same login information for other accounts, it can lead to a very large security breach for said user. It is pertinent that users practice creating strong passwords and storing them somewhere that is not easy to access or publicly accessible. Users should also turn on two-factor authentication on their social media accounts. Two-factor authentication serves as a firewall when a user's password is compromised. It is also important to not utilize common terms and themes in a user's life as a password. For example, if a user has a pet named Scout, and they post frequently as Scout, hackers may easily be able to crack said user's password if it contains the phrase 'Scout' in it. Finally, users should ensure that they are logging out of publicly shared devices to maximize their password safety.

E. Labeling and Detecting Unsafe and Triggering Content on Social Platforms

Although there are many more practices for differing kinds of scenarios with Social Media Safety, these tips cover the most common unsafe scenarios on modern social media platforms. The methods in which malicious users seek to compromise user safety evolve daily, therefore it is important to build early awareness and prepare for possibly compromising situations online. Furthermore, this analysis of safety on social media networks leads to our technical methodology, which details our solution to labeling and detecting unsafe and triggering content on social platforms.

Our program, titled 'Tweet Detector' utilizes a plethora of libraries within Python. The main library utilized to scrape data from Twitter is dubbed Twint [6], along with this we utilize the YoloV5 [18] library for object detection in media found from our scraped tweet dataset. Twitter was our first test subject in order to scrape categorized posts as it is the social media platform almost notoriously known for the spread of misinformation and political propaganda, as highlighted by Kim [13] in his study on violent political rhetoric on Twitter. Along with Twitter's infamous reputation, Twitter is also one of the most commonly scraped social media applications for machine learning projects. The YoloV5 library was utilized to test if violent information can be detected in images scraped from our dataset. However, due to the lack of training on our hashtag-specific terms, the image detection library had notable trouble detecting values that could benefit our analytical findings. Our results with the YOLOv5 library are further discussed in the Results section. In addition, we utilize a Naive Bayes classifier, a Logistic Regression classifier, and a Support Vector classifier in order to understand the shortcomings and successes of our scraping and labeling techniques. Our Naive Bayes formula is derived from the mathematical Bayes theorem, seen in Figure 2. The theorem is rewritten below in the context of this research analysis, in which our label value is renamed as Predicted Sentiment and f(i) is rewritten as a word in a tweet or text [19].

Other libraries (as shown in Figure 3) in this project include PySimpleGui [20], which was utilized with the intention to curate an interface to manually allow future research students to able to use our program to scrape labeled posts and easily detect images according to a trained dataset. Figure 4 showcases the PySimpleGui interface of the tool. Utilizing Twint's search configuration settings, our software scrapes the following data from a specified hashtag tweet: usernames, follower counts, post-like counts, image links, date posted, and hashtags. These data metrics can allow future students to perform machine learning methods and also lead to our sentiment analysis predictions from the scraped tweets. Once posts are scraped, which are specified by Twint's configuration, it outputs a CVS file of the tweet data. Figure 5 showcases the types of data scraped by the Twint library.

$$p(label f_1 f_2 \dots f_n) = \frac{p(f_1|label)p(f_2|label) * \dots * p(f_n)|label * p(label)}{(f_1 f_2 \dots f_n)}$$

Figure 2. Tweet Label Formula – Derived from Bayes Theorem

Libraries Utilized
PySimpleGui - Program Interface Twint - Twitter Data Scraping YoloV5 - Object Detection

Figure 3. Libraries Utilized in Labeling Process

Upon gathering the scraped posts, it is automatically sorted with the labelTweets function, which individually scans a tweet from the CSV file of scraped tweet information and selects or denies the tweet on a single condition: if it contains a word in a pre-established array of 'violent' terms. The final output from the labelTweets function is a compiled list of those selected tweets, related to the configured corpus settings pre-established by the user. In our sentiment analysis program, we also manually label the scraped output from the function labelTweets as negative, positive, or neutral based on if the tweet contains negative terms, according to the same corpus used in labelTweets and if the tweet contains positive terms according to our list corpus of positive terms, which is shown in Figure 6. Our negative corpus terms are shown in Figure 7. Neutral tweets are tweets that do not contain either positive or negative terms. For example, neutral tweets could be solely informational or statistical, and therefore do not contain any of the trigger corpus words. The tweets then are assigned a numerical label, where negative tweets are labeled with -1, positive tweets are labeled as 1, and neutral are labeled as 0. The code utilized to label sentiment in tweets is showcased in Figure 8.

Our code stores the sentiment scores in a list, which, is then merged with the original dataset. In order to create statistical conclusions and utilize a classifier on our data, we then must pre-process our tweets. The data promptly undergoes an

intensive cleaning process, which involves dropping integer values, punctuation, repeating characters, and removing any links. Our cleaned tweet data and sentiment values are then inserted into our sentiment analysis pipeline code, which yields the percentages of positive or negative tweets and also outputs a classification report. We have also generated a word cloud, and receiver operating characteristic (ROC) charts related to the amount of negative and positive tweets in our dataset. We then utilize the Multinomial Naive Bayes model [21], which works well with text-based data. Our data is split 70/30, with the testing size being 30 percent. Our results, which are discussed in the next section, highlight the percentages of our sentiment analysis data along with how the results of our scraping techniques result in shortcomings and successes.

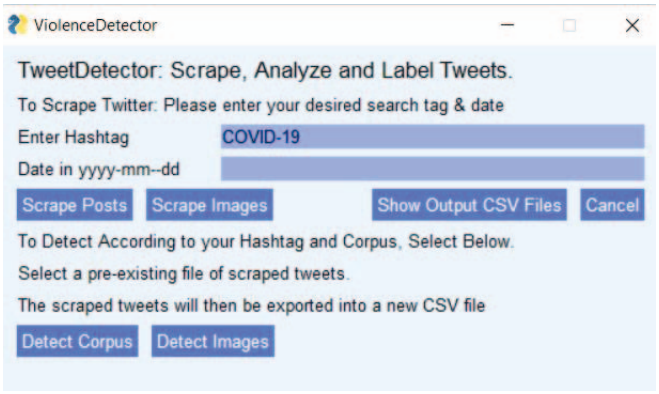

Figure 4. Tweet Detector GUI

```
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   id                  857 non-null    int64
 1   conversation_id     857 non-null    int64
 2   created_at          857 non-null    object
 3   date                857 non-null    object
 4   time                857 non-null    object
 5   timezone            857 non-null    int64
 6   user_id             857 non-null    int64
 7   username            857 non-null    object
 8   name                857 non-null    object
 9   place               0 non-null      float64
 10  tweet               857 non-null    object
 11  language            857 non-null    object
 12  urls                857 non-null    object
 13  retweets_count      857 non-null    int64
 14  hashtags            857 non-null    object
 15  link                857 non-null    object
 16  near                0 non-null      float64
 17  full_text           857 non-null    object
 18  textblob_sentiment  857 non-null    float64
```

Figure 5. Types of Data Scraped

IV. CASE STUDY

Our data consists of information directly scraped from Twitter. This process is fairly simple, however, the process of accessing specific posts requires a thorough screening from Twitter. Our project utilizes elevated Twitter API scraping privileges. Therefore, our information and research purpose had to be fully disclosed to the Twitter administrative team to allow for large queries of information to be scraped during multiple occurrences.

Overall, our scraping tool has scraped tweets from the following hashtags in Table 1. Along with these hashtags, we filtered our scraped posts according to the following short word corpuses shown in Figures 6 and 7. It is important to note that we limited our scraping terms in this case to violent or triggering scenarios. However, the way the corpus is configured in the technical process can lead to the scraping of any category of tweets selected by the user.

Hashtag Terms
"amazing","happy","wow"
"congratualtions","energetic","energized",
"enthusiastic", "exciting","dazzled"
"awesome","great","hope","hopeful"
"hopefulness","smiling","happiness","bubbly"
"beautiful","silly","gentle","encouraging"
"peace","best","bliss","blissful","blessed"
"blessing","blessings","bless","blessed"
"brave","comfy","bravery","courageous"
,"courage","courageously","courageousness"
"radiant","sweet","sweetness"
"sweetheart","sweetie","sweetheart","sweethearted","cool"
"excellent","excellence","honest","honesty"
"kind","kindness","kindhearted"
"kindly","lovely","loving"
"lovingly","lovingness","lovinghearted"
"fantastic","faith","faithful"
"faithfulness","faithfully"
"faithfullness","faithfull","faithfullly"
"fortunate","lucky","friendly","fresh","freshness"
"freshly","freshness","freshly"
"freshness","gem","freedom"
"free","freeing","glad","glee","glorious","glory"
"gracious","grace","graceful"
"gracefully","gracefulness"
"gift","genuine","generous","generosity","generousness"
"gleeful","help","helpful"
"helpfulness","helping","honor"
"honorable","honour","honourable","honourably","honouring"
"healthy","healthy","healthful"
"healthfully","humane","miraculous"
"miracle","miracles","miraculousness","miraculously"
"genius","good","goodness","goodly"
"harmony","harmonious","harmoniously"
"friendly","friend","friendship","trust","safe","safety"

Figure 6. Corpus of Positive Terminology

Corpus Terms
"Covid-19", "COVID", "hate", "ew"
"gross", "sick", "disease", "sneeze"
"cough", "pandemic", "epidemic", "new"
"fear", "toll", "spreading", "declare"
"infect", "Wuhan", "China", "COVID-19"
"Impact", "fight", "patient", "death"
"concern", "epidemic", "strain", "symptom"
"Spreading", "scary", "scared", "shooting"
"school shooting", "2019", "infect"
"deadly", "outbreak", "respiratory", "quarantine"
"virus", "infect", "not", "SARS"
"PPE", "disinfect", "isolation", "self-isolation"
"lockdown", "sanitizer","sanitize", "evacuee"
"distancing", "impeachment", "airstrike", "bombed"
"strikes", "riot", "crowds", "non-essential", "corona"
"militia", "evacuate", "war", "plague"
"emergency", "infection", "deportation", "swine"
"punishment", "nuisance", "shoot", "lowlife"
"drugs", "cocaine", "meth", "marijuana"
"fighting", "fights", "violent", "violence"
"brutality", "cruelty", "bloodshed", "deport"
"wanker", "beating", "propaganda", "fake"
"clash", "murder", "foul play", "blood"
"attack", "beating", "jumped", "rape"
"coercion", "assault", "crimes", "crime"
"mortality", "fatality", "fatalities", "n-CoV"

Figure 7. Corpus of Negative Terms

To conclude, our total dataset of scraped tweets then later analyzed for sentiment in our results is a compilation of 500 tweets related to the topics and terms highlighted in Figures 3, 6 and 7. The time frame of scraped information begins as early as August 2019. The variety of scraped information is not exclusive specifically to 2019. The data scraped ranges from 2019 until the present day. The scraping is also not sequential; it is scraped based on the hierarchy of popularity within the specific hashtag on Twitter. The data utilized in the sentiment analysis process is also a variety of data scraped according to all of the hashtags aligned in the chart below.

```
lst = []
for i in range(len(df_labelled.index)):
    if any(x in df_labelled['tweet'][i] for x in negcorpus):
        lst.append(-1)
    if any(x in df_labelled['tweet'][i] for x in poscorpus):
        lst.append(1)
    else:
        lst.append(0)
```

Figure 8. Sentiment Analysis Assignment Code

Hashtag Terms
'COVID-19'
'School Shootings'
'Election'
'Pandemic'
'War'

Figure 9. Hashtag Terms Utilized to Scrape Tweets

Social media platforms such as Twitter, Facebook, and Instagram (social media hosts), although created with positive intent, can often allow for the spread and amplification of very negative and offensive discourses. This issue was a point of interest for researchers Nobal B. Niraula of Nowa Lab, Saurab Dulal of The University of Memphis, and Diwa Koirala of Nowa Lab in their paper "Offensive Language Detection in Nepali Social Media" published in 2021. This case study will focus on the experimentation done by these researchers and the results they obtained.

In their paper, Niraula, Dulal, and Koirala highlight the lack of resources for language technologies in languages like Nepali, coined a low-resource language, as opposed to a resource-rich language like English or German. Aiming to characterize the offensive language present in Nepali social media, researchers present experiments using supervised machine learning to contribute data and the first baseline approaches of offensive language detection in this low resource language.

Throughout their paper the researchers refer to foul language as typically consisting of racial hate speech, personal attacks, and sexual harassment. Addressing this issue is important because of the large volume of comments or posts on social media platforms that may contain toxic tones and are acutely insulting or harmful to other users. By eliminating foul language, the environment of social platforms can remain positive while maintaining healthy discussions and enhancing the security of the users. Inspired by prior works, the researchers have several key contributions to this field of study. Firstly, they characterize the offensive, toxic, and foul language commonly found in Nepali social media. This allowed them to release a human-labeled data set for offensive language detection in Nepali social media, available at: https://github.com/nowalab/offensivenepali. This differs from our approach at gathering Tweets. Where the researchers manually gathered tweets according to the foul-language targets, we were able to gather targeted tweets using the Twitter scraper. The researchers then developed novel preprocessing approaches for Nepali social media text. Finally, the culmination of their research led them to provide baseline models for coarse-grained and fine-grained classification of offensive language in Nepali.

V. RESULTS

To begin, our results highlighted in our Naive Bayes classifier report (see Table 1) are mainly affected by our scraping technique, which was mostly handled by Twint. Due to the processing power needed to scrape large amounts of data, there were instances where Twint would have difficulty scraping more significant amounts of data. Therefore, our resulting statistics may be affected by the small size of our initially scraped dataset which, in its final state, contains a range of 500-600 tweets. Along with this, through the formatting of the program, and due to how Twint searches for

tweets, it was difficult to find tweets that had an intersection of our highlighted key hashtag search terms in our Hashtag Terms chart (Figure 9). The prediction scores for our Multi Nominal Naive Bayes classifier are showcased in Table 1. The scores highlighted reflect how many tweets were correctly classified. The scores yielded by the classifier are surprisingly much lower than expected, however, there is an outlier of 81 percent under the f-1 score. These numerical results may be a result of the lack of strength in our dataset. A more thorough and extensive cleaning process with visualization included may assist in yielding higher scores.

Another factor contributing to the classifier's poor performance may be the size of our dataset. Once scraped, our dataset contains about 500-600 tweets, which are continuously narrowed down by the labeling process. Many tweets that could contain information that would benefit our analysis can be lost in the labeling process. A possible fix may be an in-depth fine-tuning of the tweet selection process and the sentiment labeling process or fully switching our manual automated sentiment labeling to then having our tweets labeled by humans.

Table 1. Naive Bayes Classifier Scores

Precision	Recall	F1-Score	Support
0.22	0.53	0.31	19
0.88	0.75	0.81	209
0.40	0.47	0.43	30
0.50	0.58	0.70	258
0.78	0.70	0.73	258

Table 2. Support Vector Machine Classifier Scores

Precision	Recall	F1-Score	Support
0.62	0.17	0.27	29
0.81	0.98	0.89	200
0.67	0.21	0.32	29
0.70	0.45	0.80	258
0.78	0.80	0.75	258

Table 3. Logistic Regression Classifier Scores

Precision	Recall	F1-Score	Support
0.53	0.28	0.36	29
0.83	0.95	0.89	200
0.57	0.28	0.37	29
0.64	0.50	0.80	258
0.77	0.80	0.77	258

Next, we ran our Logistic Regression model (see Table 3) on our sentiment-labeled dataset, which resulted in mid to higher scores. This model performed significantly better than the Naive Bayes Model, which is expected as Logistic Regression classifiers are mainly used when the dependent variable is categorical, which is the case for our dataset; it categorizes our data according to our three categories, negative, positive, and neutral. Finally, our Support Vector Machine classifier (see Table 2) also outperforms our Naive Bayes classifier and produces high recall and precision score values for the positive and negative sentiment values, ranging from 95 to 98. This can be attributed to a Support Vector Machine classifier producing better results on a linear dataset. Therefore, it is expected for the third category, in this example, the first row in our data seen in Table 2, to score lower. Along with these difficulties, the YOLOv5 object detection library also struggled to manually label each image from the links in CSV files, as the model was not thoroughly trained according to our chart of hashtag search terms. The data found by YOLOv5 were things like 'man', 'tie', and 'sandwich'. While it is interesting that these items were recognized in the media from our dataset, it ultimately did not add to our goal of accurately detecting violent images from our scraped social media posts.

Furthermore, the data yielded from the YOLOv5 image analysis did not accurately detect signs of concepts highlighted in our hashtag search terms, such as 'COVID19', 'War', 'School Shooting', or 'Pandemic'. Our Word Cloud, shown in Figure 10 is generated based on the frequency of terms found in our tweet dataset. The most prominent terms in the word cloud are 'Covid-19', 'coronavirus', 'vaccine', 'death', 'people', and 'Pfizer'. Terms related to COVID-19 are notably the most prominent in the word cloud due to the time frame from which our tweets were scraped, which began in 2019. The years 2019-2021 were defined by the COVID-19 pandemic, hence resulting in the influx of terms related to COVID-19 in the word cloud. Therefore, through our numerical findings resulting from sentiment analysis of our collected tweets, it is evident that a majority of the tweets classified and showcased in the word cloud are negative and therefore, align with our selection algorithm of corpus selection terms.

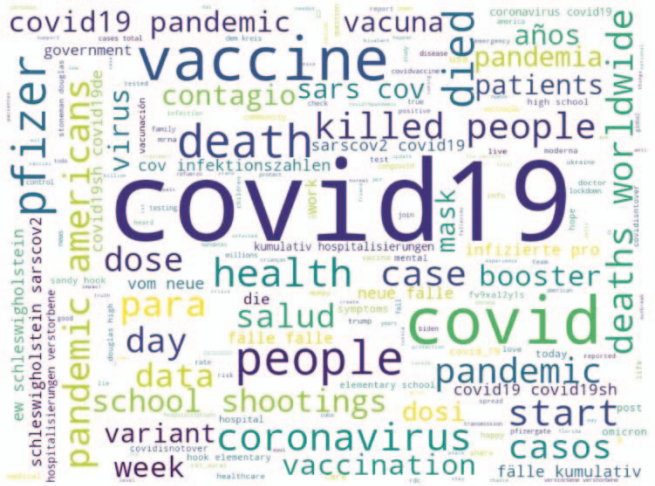

Figure 10. Negative Word Cloud

VI. THREATS TO VALIDITY

Some possible threats to our validity include utilizing a library in order to scrape tweets, which may lead to a limited

amount of tweet collection instead of a manual hand-picked method of scraping. Utilizing terms related to COVID-19, along with this our dataset size may lead to invalidity. Limiting our hashtag term to scrape data from Twitter is also a threat to the validity of our search, as there are many terms that could be utilized to search. We chose a select amount of hashtags in order to encompass a short yet important variety of tweets that have been hot topics in recent years in the United States specifically.

VII. CONCLUSION

Our research has dealt with social media safety practices, flagging sensitive posts, and effective content monitoring. In this paper, an examination was made of the most impactful hazards and mishaps that arise from poor social media content monitoring, incorrect flagging, public safety, and the spread of misinformation. In addition, this paper discussed the shortcomings of content monitoring tools currently on the market and how they can be improved. Through the synthesis and culmination of exploring social media safety practices, we have curated an application to label social media posts according to a pre-established corpus of violent phrases.

Thus far, our research has provided results that showcase the sentiment shown in the scraped tweets and we now seek to improve our tool with the capability of multiple platforms, along with the capability to scrape more advanced batches of data. In the future, we seek to scrape Facebook, Twitter, Instagram, and possibly, Reddit. In order to facilitate this growth on the software, it would take a significant amount of time to debug and complete the coding and implementation of libraries utilized to scrape those specific social media platforms. The versatility of our software allows users to search and scrape for terms and images customized to their preference, therefore as this project evolves, while its focus in this context was to monitor for more modern sensitive tags on Twitter, this project can evolve alongside the social state of social media platforms to a user's settings.

ACKNOWLEDGMENT

This work was supported by the USA National Science Foundation under Grant 1757828, Grant 1822137, and Grant 2050869.

REFERENCES

[1] M. Iqbal, "Twitter revenue and usage-statistics." Available at https://www.businessofapps.com/data/twitter-statistics/.

[2] D. Burrows, "Investigative committee on the robb elementary school(uvalde) shooting - interim report."

[3] B. P. et al., "Cyber bullying awareness:-major cause of mental health problems amond adolescent of selected school, chhotaudepur."

[4] A. K. e. a. Ramez Kouzy, Joseph Abi Jaoude, "Coronavirus goes viral: Quantifying the covid-19 misinformation epidemic on twitter." Cureus 12.3.

[5] Tweepy, "Tweepy." Available at https://docs.tweepy.org/en/stable/. [6] Twint, "Twint." Available at https://github.com/twintproject/twint.

[6] D. Lauer, "Facebook's ethical failures are not accidental; they are part of the business model." AI and Ethics 1.4.

[7] S.-M. W. Group, "Countering false information on social media in disasters and emergencies." Available at https://www.dhs.gov/publication/st-frg-countering-false-informationsocial-media-disasters-and-emergencies (2018/03).

[8] L. O'Leary, "Why expensive social media monitoring has failed to protect schools." Available at

[9] https://slate.com/technology/2022/06/social-media-monitoringsoftware-schools-safety.html (2022/06/04).

[10] R. W. Candice Lanius and W. MacKenzie, "Use of bot and content flags to limit the spread of misinformation among social networks: a behavior and attitude survey." Social Network Analysis and Mining 12.1.

[11] J. Y. Wallace Chipidza, "The effectiveness of flagging content belonging to prominent individuals: The case of donald trump on twitter." Journal of the Association for Information Sciecne and Technology 73.11.

[12] T. Kim, "Violent political rhetoric on twitter." Political Science Research and Methods.

[13] B. V. Paul Rottger, "Hatecheck: Functional tests for hate speech detection models." arXiv preprint arXiv:2012.15606.

[14] S. D. Nobal B. Niraula, "Offensive language detection in nepali social media." Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021).

[15] D. S. John Allen Hendricks, "The social media election of 2016." The 2016 US presidential campaign.

[16] S. K. Ishfaq Majid, "Social media and security: How to ensure safe social networking." Social media and security: how to ensure safe social networking. International Journal of Humanities and Education Research 1.1.

[17] G. Jocher, "Yolov5." Available at https://docs.ultralytics.com/.

[18] N. Jain, "Sentiment analysis using naïve-bayes." Available-at https://www.enjoyalgorithms.com/blog/sentiment-analysis-usingnaive-bayes).

[19] M.Barnett,"Pysimplegui." Available at https://www.pysimplegui.org/en/latest/.

[20] F. P. et al., "Scikit-learn: Machine learning in python." Available at http://jmlr.org/papers/v12/pedregosa11a.html.

[21] A. Cipriano, "Third party student surveillance: Is monitoring student speech going toprevent the next school shooting?." Rutgers UL Rev. 72.