Social Network Analysis and Mining (2023) 13:51
https://doi.org/10.1007/s13278-023-01053-4

ORIGINAL PAPER

An NLP-assisted Bayesian time-series analysis for prevalence of Twitter cyberbullying during the COVID-19 pandemic

Christopher Perez[1] · Sayar Karmakar[1]

Received: 22 August 2022 / Revised: 19 February 2023 / Accepted: 24 February 2023 / Published online: 15 March 2023

Abstract
COVID-19 has brought about many changes in social dynamics. Stay-at-home orders and disruptions in school teaching can influence bullying behavior in-person and online, both of which leading to negative outcomes in victims. To study cyberbullying specifically, 1 million tweets containing keywords associated with abuse were collected from the beginning of 2019 to the end of 2021 with the Twitter API search endpoint. A natural language processing model pre-trained on a Twitter corpus generated probabilities for the tweets being offensive and hateful. To overcome limitations of sampling, data were also collected using the count endpoint. The fraction of tweets from a given daily sample marked as abusive is multiplied to the number reported by the count endpoint. Once these adjusted counts are assembled, a Bayesian autoregressive Poisson model allows one to study the mean trend and lag functions of the data and how they vary over time. The results reveal strong weekly and yearly seasonality in hateful speech but with slight differences across years that may be attributed to COVID-19.

Keywords Time series · Cyberbullying · Bayesian estimation · COVID-19 · Twitter · NLP

1 Introduction

Technological developments throughout history have fundamentally changed how people communicate and interact with one another. With new successes come new challenges, as the rapid proliferation of the internet has led to a phenomenon known as cyberbullying. Many questions may arise, such as how cyberbullying can propagate and how it interacts with global crises such as the COVID-19 pandemic. In this paper, we combine natural language processing and Bayesian time-series analysis methods to provide a systematic assessment of the trends of cyberbullying for a span of three years. We begin this exposition with an informal description of cyberbullying before describing the connection with COVID-19 and Twitter.

✉ Sayar Karmakar
sayarkarmakar@ufl.edu

Christopher Perez
cperez7@ufl.edu

1 Department of Statistics, University of Florida, Gainesville, FL 32601, USA

1.1 Describing cyberbullying

One definition for cyberbullying is "*An aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself*" (Smith et al. 2008). And with increases in computer use, the possibility for cyberbullying grows.

An immediate question is what sets cyberbullying and traditional in-person bullying apart. These differences are paramount when considering a theoretical approach to studying cyberbullying, as models made with traditional bullying in mind may not tell the two apart (Barlett 2017). One major point is that cyberbullies can maintain anonymity online which makes it difficult to locate perpetrators (Bonanno and Hymel 2013). Due to the reach of social networks, cyberbullying may also persist far beyond the reaches of normal bullying and can proliferate to large swaths of people, often attaining viral status (Aboujaoude et al. 2015). Cyberbullying is a perpetual phenomenon that constantly places stress on the victim.

Cyberbullying has been studied to be a cause of many negative outcomes in victims. A meta-analysis conducted in the topic reveals correlations with low self-esteem, depression, and drug abuse (Kowalski et al. 2014). Many episodes of attempted suicide and self-harm have been directly attributed to cyberbullying (Kwan et al. 2020). Interestingly, according to surveys, only a small percentage of cyberbullying victims are not bullied in a traditional, in-person manner (Olweus and Limber 2018). Regardless, the many negative outcomes and pervasiveness of cyberbullying has led to some to suggest it is a serious public health threat, and its danger can only grow with increased mobile device and social media usage (Aboujaoude et al. 2015).

1.2 COVID-19 and cyberbullying

The proliferation of COVID-19 has significantly changed the way many people live. Lockdowns have been put in place to curb the spread, but not without consequences. As humans thrive in social situations, the isolation of many from the day-to-day affairs has led some to posit that people's mental health will worsen, noting consequences such as maladaptive behaviors, loneliness, and depression (Talevi et al. 2020). Direct research shows that quarantines and self-isolation are linked with higher prevalence of issues like depression and insomnia (Wang et al. 2021a). When comparing time periods before and after the COVID-19 pandemic, Barlett et al. (2021a) suggest that important components of their cyberbullying model, such as cyberbullying attitude, cyberbullying behavior, and belief in irrelevance of muscularity in online bullying, have significantly changed between these time points. It is also observed that cyberbullying is correlated with COVID-19 experiences (Barlett et al. 2021b).

The interaction of COVID-19 and cyberbullying in academic settings is a topic of great interest. As universities and schools around the world shifted to online instruction to deter spread of the virus, approximately 1.5 billion students have had their education interrupted (Bozkurt et al. 2020). The effects on isolation on university students is an important matter, as these groups show high proportions of common mental disorders. Such literature reveals these groups were associated with more frequent internet use and may thus have increased probability of being involved in cyberbullying (Mota et al. 2021). Quantitative analysis of such groups in India (Jain et al. 2020) have shown that 80% of those between 17 and 18 years old were bullied during the pandemic, and 79% of those experiencing traditional bullying before the pandemic were cyberbullied during pandemic, corroborating the notion that victims of traditional bullying are also victims of cyberbullying. Other categories of victims show an increase in percentage of those cyberbullied from before to during the pandemic as well.

1.3 Using Twitter to understand social phenomena and cyberbullying

As in the aforementioned studies, much research in cyberbullying involves the use of survey data. However, if one relies on a responder's willingness to self-report, then that leaves the door open to problems such as responder bias and invalid responders. In a study on adolescent regarding risk behavior, it is found that responders who purposefully answer wrongfully "showed" higher rates of such behavior, such as alcohol and drug consumption (Cornell et al. 2012). Such a happening may be present in bullying surveys as well, which can be exacerbated by sample-size limitations imposed by cost and time. Additionally, if a researcher was interested in studying the impact of COVID-19 on cyberbullying, they ideally have to collect data before the pandemic began, as done in previous research (Barlett et al. 2021a). Since the magnitude and impact of global crises are sometimes unforeseeable, it can be difficult to know when to start such a longitudinal study.

Social media can serve as rich data source that remedies some of the issues that affect survey collection. One major advantage is the scope of social media. Twitter, the choice for our study, had approximately 186 million users, 36 million of which from the USA [1] in 2020. Furthermore, the availability of web scraping technology and even an official API allow individuals to use Twitter's public archive of tweets dating back to 2006. This software is often free to use, making large-scale studies much more affordable. Information harvested from these tweets can be used for many purposes, such as monitoring disease spread and forecasting elections (Signorini et al. 2011; Tumasjan et al. 2010). This data source enables us to perform a study reminiscent of longitudinal study, with the ability to collect previous years of data without the associated cost of maintaining a large-scale study for many years.

However, one may question the efficacy of using social media such as Twitter as a data source to understand social phenomena. For example, Tumasjan et al. (2010) showed that using Twitter traffic to predict vote share in German errors resulted in very low prediction error. Signorini et al. (2011) demonstrate a correspondence between Twitter data and the H1N1 at the overall national level, as well as smaller geographic regions. Based on real-time data, their estimates could be produced earlier than regular health reports. While these present advantages, one must also consider the possible limitations. For instance, predicting elections using Twitter faces two major issues, one being that sampling data from social media does not match the sophistication of more developed polling processes, and that spam, propagandists, and fake accounts can easily manipulate data (Gayo-Avello

[1] https://www.businessofapps.com/data/twitter-statistics/.

et al. 2011). In the influenza study conducted by Signorini et al. (2011), the researchers faced limitations in the lack of uniformity of Twitter usage by different locales and in different time periods. They also could not generalize to a population beyond some form of Twitter population. These limitations are also present in the current study. Brief use of the search endpoint for an arbitrary query may show a few spam or bot posts, heightening the importance of filtering these posts out. Additionally, since Twitter's sampling algorithm is unknown, it is difficult to identify the exact population the study's results can be generalized to. There are methods to reduce this uncertainty by using geo-tagged tweets, as retrieving tweets of this nature results in a more complete sample (Morstatter et al. 2013).

Despite limitations, there are a variety of points to make in justifying Twitter as a data source to study cyberbullying. McHugh et al. (2019) suggests that Twitter is a hotbed for "intentionally aggressive, harmful communication." They explain results from surveys showing that about 70% of college students use Twitter, and the amount of cyberbullying on Twitter was gauged to be higher than other platforms such as Facebook and Instagram. Additionally, the official Twitter API [2] allows users to perform a variety of queries and searches on its public archive. With this technology, we can collect swaths of tweets satisfying certain specifications, such as containing particular keywords or hashtags, in order to study the cyberbullying problem.

To analyze the tweets retrieved from Twitter in the first place, many researchers turn to natural language processing (NLP). The usage of NLP in this study is similar to that of previous work done regarding Twitter cyberbullying and COVID-19 (Babvey et al. 2021). The study by Babvey et al. (2021) uses NLP as a pre-processing step to filter out tweets that have low probability of being abusive speech, and then examine the difference in number of such tweets before and after a fixed time period. In this study, we use NLP to perform a similar pre-processing step, but instead of comparing a fixed time point, we study each day from 2019 to 2021. Few studies have attempted to study cyberbullying in response to COVID-19 from a continuous perspective. Researchers using Google Trends time-series data (Bacher-Hicks et al. 2022) demonstrate that cyberbullying was actually disrupted by COVID-19, while others, using Twitter data (Karmakar and Das 2020), claim that it led to an increase. We seek to address these studies by broadening the time frame of analysis and using more thorough methods, namely NLP, to obtain better samples to study cyberbullying patterns with.

[2] https://developer.twitter.com/en/docs/twitter-api.

1.4 Considerations and assumptions for cyberbullying detection

As we seek to understand the general volume of cyberbullying over several years of time, we are forced to make certain assumptions that allow us to work with the appropriate type of data. In light of our choice of a pre-trained NLP model, we assume that two metrics, hatefulness and offensiveness of a tweet, are associated with cyberbullying. However, detecting cyberbullying in an extremely accurate fashion may be a more difficult task than this assumption may imply. Many detection models in literature employ more advanced considerations to detect cyberbullying events such as images, location, and a given user's profile and comment history, which may contain vital information to predict cyberbullying behavior Cheng et al. (2019a); Dadvar et al. (2013). Other models involve hierarchical attention networks to make use of the inherent structure of social media as additional context for predicting (Cheng et al. 2019b), or constructing graphs composed of sender and receiver nodes to mimic cyberbullying interactions (Huang et al. 2014). In our study, the NLP model of choice does not necessarily make use of this more complicated information, and thus its accuracy may suffer relative to more state-of-the-art methods. Regardless, we choose this model for its off-the-shelf accessibility and hence tailor our analysis to what the model is capable of delivering.

1.5 Summarizing our contribution

For our study, we collected 1,004,466 tweets from January 1, 2019, to December 31, 2021, with the Twitter API's search endpoint based on keywords used in previous cyberbullying and abusive speech studies (Wiegand et al. 2019; Nand et al. 2016; Cortis and Handschuh 2015). To clean the data, a pre-trained NLP model tuned to classification of offensive and hateful tweets models the probability of a given tweet being offensive or hateful Barbieri et al. (2020). Because of irregularities in the search endpoint's returned sample sizes, the count endpoint is used with the same keywords as before to obtain a more consistent and comprehensive count of all tweets that the search API could have picked from. The tweets from the search endpoint are then filtered using certain probability thresholds to select relevant tweets. Then, for each day in the study, the fraction of relevant tweets out its respective sample is calculated. This fraction is multiplied by the number reported by the count endpoint, generating a time series that takes into account the different proportions of abusive content on each day while addressing issues with the search endpoint's sampling procedure. This data collection and filtering is described in Sect. 3.

Visual analysis of the data collected is in Sect. 4 and motivates our usage of a Bayesian time-series model in

Sect. 5. The results demonstrate that the tweets likely to be hateful speech exhibit strong weekly and year seasonality, which is not as evident in the unfiltered data. Patterns such as two distinct increases in mean trend during the first and second halves of the year remain constant throughout 2019 through 2021, but with slight differences. A time series of new COVID-19 cases is fit to the same model and similarities between this model and the Twitter data are discussed. Further, the effect of the pandemic on these trends, if at all, is to decrease the scale of potential cyberbullying tweets, as well as widen the second peak of the year. The proportion of these potential cyberbullying tweets from their respective daily samples also seems to roughly decrease over time. The results can help guide developers of Twitter or other social media to ramp up mitigation technology at the appropriate time by taking the strong seasonal behavior into account, and overall, introduce novel ways of examining a familiar problem.

The work is concluded by a discussion of the issues and limitations of the study, the implications of the results found, and avenues for future research.

2 Related work on internet data, cyberbullying, and COVID-19

This study is motivated by the findings of Karmakar and Das (2020), which employed a Bayesian, time-varying linear Poisson autoregressive model to tweet counts containing keywords related to cyberbullying. Such an analysis was the first of its kind in this field. Their study, confined to the first half of 2020, concluded a rise in mean trend from March to April similar to that of COVID-19 cases and that the first lag accounted for most of the correlation. However, there is criticism to be made in that preliminary analysis. The choice of keywords along with lack of text analysis confined the results to cyberbullying discourse rather than actual cyberbullying events. Cyberbullying attacks may precipitate awareness and discourse, but they do not follow the same time series. While their work collected data using web scraping, the current study pulls data directly from the Twitter API.

Another work by Bacher-Hicks et al. (2022), instead of using Twitter data, opts to use Google Trends, a site that provides time-series data for search intensities of search terms.[3] This work studies search intensity of cyberbullying and bullying. Like the previous study, they do not depict the frequency of cyberbullying events exactly. Further, since the main model in our study is based on a Poisson distribution, which requires count data, we cannot use the TVBARC model on Google Trends data in attempt to replicate the study.

However, consider the two following similar findings from these works. The model fit by Karmakar and Das (2020) reveals an increase in mean trend from around March to May of 2020, while the model constructed by Bacher-Hicks et al. (2022) shows that the deviation from predicted log search intensity increases roughly around the same time frame. One important distinction, though, is that the Bacher-Hicks study includes data from before January 2020 and slightly after. Bacher-Hicks claims, further, that this rise in the March-May period was just an increase back to levels before the onset of the pandemic, and that COVID-19 had disrupted cyberbullying. In the absence of a wider time frame in Karmakar's paper, one may conclude an increase in cyberbullying discourse on Twitter, but this may very well suggest a return to pre-pandemic levels as Bacher-Hicks describes. Again, one cannot say that this necessarily extends to cyberbullying events, but how it becomes a trending topic over time. Furthermore, social media and search engines are used with different motives in mind, which may result in discrepancies in findings (Li et al. 2021).

The work done by Babvey et al. (2021) motivates our decision to employ an NLP model in an attempt to retain true cyberbullying events. They query Twitter for keywords associated with abusive speech and then run a machine learning model to discard tweets that are likely not abusive. By using such methods, they are able to have more confidence that their data can represent actual cyberbullying events. They compare two sets of data collected before and after March 2020 to gauge the effect of COVID-19 and the associated interactions with cyberbullying. Their results show an increase in prevalence of abusive and hateful tweets once the pandemic-era lockdowns began.

Usage of NLP methods can help avoid the issue of covering cyberbullying discourse rather than potential cyberbullying events. The use of a larger time frame along with a more continuous time-series approach allows one to see whether COVID-19 has a sustained effect on Twitter cyberbullying, or if previous findings may have been coincidences or one-time occurrences. Now, the data collection and NLP filtering procedure are discussed.

3 Data collection and cleaning

A straight-forward way to access to Twitter data is by using the official Twitter API. To make the most use of the Twitter API, we were provided with an Academic Research License, granting us features such as the full-archive search. This is critical to our research, as it lets us search many years' worth of data quite easily. We used the R programming language to interact with the API.

[3] https://trends.google.com/.

3.1 A foreword on the Twitter API

Unfortunately, the algorithm used by the Twitter API to sample tweets is unknown. As studied by Thelwall (2015), the search endpoint may not be comprehensive. However, the tweets that were not retrieved by its sampling procedure are more likely to be spam. Other works (Morstatter et al. 2013) point out that the sampling tweets may result in decreased accuracy (as compared to alternative, costly methods to acquire every single tweet), but interestingly, the sampling algorithm recovers a higher proportion of tweets that are geo-tagged.

The nature of time-series analysis emphasizes these sampling issues. The subset of tweets taken from all matching tweets may not be a fixed percent, so the relative sizes between daily counts is not preserved. Simple repetitions of identical requests may sometimes return more tweets seemingly at random. Comparison with the number reported by the count endpoint is an enticing option for a couple reasons. The count endpoint returns much more consistent results through runs, and since it does not have to go through additional compliance that the search endpoint does, the data returned may be more complete.[4] To this end, we take the percentage of relevant tweets from a given daily sample and multiply that percentage to the number reported by the count endpoint. Selection of relevant tweets is described below.

3.2 Collection procedure

To collect a representative dataset, we first sample tweets using the search endpoint to access their textual content. A list of keywords must be assembled to query for in both the search and count endpoints. As a starting point, we reference a lexicon provided by Wiegand et al. (2019). It contains a list of words each with a score rating its abusiveness according to their trained model. One may notice that identical words appear more than once (as a verb and as a noun, for instance), and hence we only keep the highest score and discard the other entries. From this adjusted list, the 100 highest ranking words were taken.

For comparison purposes, we reference two similar studies of cyberbullying analysis (Nand et al. 2016; Cortis and Handschuh 2015). Their lists of keywords are different and are only composed of about 25 and 10 words, respectively. To test the efficacy of our 100 keywords, we sample tweets in January 2020 using our original list of words, specifying no retweets, written in English, and based in the USA. We then count the number of tweets each keyword in our list appeared in, including keywords in the other studies that were not in our original list. Based on numbers of tweet occurrences, we again take the 100 highest performing words.

With this new list that combines information from the aforementioned studies, we use the search endpoint to query the entirety of 2019, 2020, and 2021. Like before, we specify no retweets, tweets written in English, and tweets from the US, but this time we also specify no promotional tweets. Each day contains anywhere between 400 and a couple thousand tweets with their textual content. In total, we collected 1,004,466 tweets. The same query is also used for the count endpoint to collect daily counts. Once the full data set is assembled into a data frame, the results are written into CSV files for storage.

3.3 A pre-trained NLP model for pre-processing data

As discussed before, in order to assemble a time series of cyberbullying events, certain assumptions may be made. Using user profile information in a large-scale time-series context may prove to be difficult, and thus we choose to make assumptions that make the processing analysis more straightforward. Encouraged by our choice of NLP model (Barbieri et al. 2020), we use two potential proxies for cyberbullying, being the hatefulness and offensiveness of textual content of tweets. By denoting tweets that meet a certain threshold as those most likely to be cyberbullying events, we can easily construct a time series to perform ensuing analyses. The provision of NLP as a filtering mechanism is an improvement that has great potential in cutting down spam and irrelevant tweets, whereas previous work only relied on the number of matching tweets based on keywords Karmakar and Das (2020). The NLP model in question is available freely for use on HuggingFace [5] .[6]

To use these models, we work with the Python language in Google Colaboratory ,[7] which provides a high performance cloud computing environment and greatly simplifies set-up of packages and other dependencies. Instructions for basic setup to use HuggingFace are found on the website.[8] Template code for using the models is found on the model pages. While there is a text pre-processing step that is part of the template code, it does not contain the additional step of removing line-breaks that the authors of the model, Barbieri et al. (2020), used in their own analysis, so it was added to the text pre-processing function. We also converted all text to lowercase to handle erratic capitalization and removed duplicate white space. The fitting procedure was generalized to collections of text using a for loop.

[4] https://developer.twitter.com/en/docs/twitter-api/tweets/counts/introduction.

[5] https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive.

[6] https://huggingface.co/cardiffnlp/twitter-roberta-base-hate.

[7] https://research.google.com/colaboratory/.

[8] https://huggingface.co/course/chapter0/1?fw=pt.

Using two of their models, we can produce the probability of tweet being offensive and the probability of a tweet being hateful. To filter out irrelevant data, we must properly select thresholds for each of these scores. Motivating this discussion, we first observe a few example Tweets (identifying information is censored), where H denotes the hatefulness score and O the offensiveness score according to the NLP model.

@USER @USER Then Dr. Fauci and others should speak out every single day and defend the health and safety of US Citizens. (H: 0.07, O: 0.10)

@USER Thanks. Now she needs to make it through intensive care tonight (H: 0.03, O: 0.04)

@USER @USER @USER @USER @USER @USER IF ANY of these african migrants have ebola-can't that be spread thru water? or am I wrong? I thought ebola was spread w/fluids-any body fluids-can anyone inform me?thanks in advance. (H: 0.62, O: 0.16)

I feel like we all went to school with a bitch like this and wanted to shove her down the mf stairs [URL] (H: 0.86, O: 0.92)

One particular aspect the model is good at is removing Tweets that are clearly not hateful or offensive, and are hence very unlikely to be cyberbullying, so filtering out Tweets with lower scores on these metrics may help retain better tweets. However, while not depicted, randomly sampling tweets with a high offensiveness/hatefulness rating returned many tweets containing African-American English Vernacular, where the content of the tweet is not necessarily a cyberbullying event. This is part of a larger issue of systemic racial bias in a large swath of hate speech and abusive language datasets (Davidson et al. 2019). Unfortunately, not much can be done about this in this context without further complicating the study, potentially going out of scope of the original intentions. At the very least, we can be confident that many tweets not indicative of cyberbullying will be removed after filtering, so the dataset will be relatively more representative even with the aforementioned issues.

3.4 Subsetting truly offensive or hateful tweets

Due to our assumption of hatefulness and offensiveness as indicators for cyberbullying, we have two different metrics to work with to help determine whether a tweet may be recorded as a cyberbullying event. We do this by selecting thresholds for each of these metrics to filter out irrelevant data. To begin, we first explore the distribution of the scores themselves. In Fig. 1, one may notice that the scores of offensiveness are bimodal with peaks 0–1, with fewer tweets near the center. The high presence of tweets near 1 is likely related to the nature of the query. In the construction of the model by Barbieri et al. (2020), they used a dataset created by a different group for a similar task. The creators of said

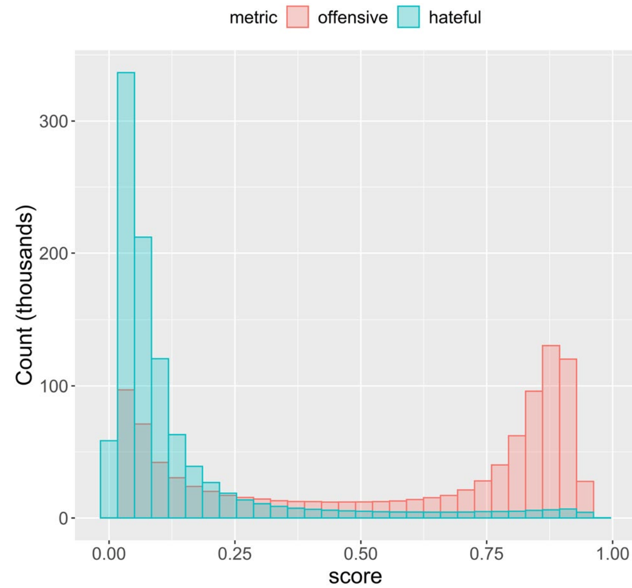

Fig. 1 Modeled probabilities of offensiveness/hatefulness of tweets according to the NLP model

dataset, Zampieri et al. (2019), describe a tweet as offensive if "it contains any form of non-acceptable language (profanity) or a targeted offense... This category includes insults, threads, and posts containing profane language or swear words." Thus, if our query contains many profane keywords, we are likely to see many tweets with a high predicted probability for offensiveness.

On the other hand, hatefulness scores are mostly lower than 0.25, and much fewer tweets have scores beyond. To reiterate, the model was fine-tuned to detect hatefulness against two target groups, being women and immigrants. The lower presence of high-probability tweets can be partially explained by queried words, as they lack many words that are explicitly targeting women or immigrants, such as those used by the authors of the dataset (Basile et al. 2019).

There are some tweets which the model failed to fit. These tweets all contained copious amounts of emojis which caused issues with the model's tokenizer. Out of the 1,004,466 tweets in this study, only seven failed to process. Each of these tweets occurred on a different day, making it exceedingly unlikely for them to affect analysis.

Now we observe the effects of subsetting tweets with scores strictly greater than a certain probability threshold, getting the percentage of those tweets from their respective daily sample, then multiplying that percent to the number reported by the count endpoint. Different thresholds are used, and the series is superimposed by a GAM smoother. The results for this process, varying offensiveness while fixing hatefulness, are displayed in Fig. 3. The local maxima are kept intact until accepting only larger scores of greater than 0.8. Likewise, for filtering on hatefulness in Fig. 4, the data are much more sensitive to applying greater thresholds. When applying a threshold of 0.05 on hatefulness, the values in January 2019 were around 60,000, and then hovered around 45,000 until 2021. A similar pattern is exhibited when thresholding at the much greater value of 0.5 on offensiveness. These results are a consequence of the distribution of overall scores as shown in Fig. 1, with offensiveness scores being bimodal near 0 and 1, while hatefulness clusters near 0 and tapers off rapidly.

Now we are tasked to choose a threshold for our dataset, then feed this data into the TVBARC model. In some problems, 0.5 may be used, but we know from reality that the occurrence of cyberbullying events is not as common as that would suggest. However, given that our query contains keywords associated with abusive behavior, it is possible that the proportion of cyberbullying events among the collected tweets will be higher. Other considerations include the NLP models' performance on their associated test data. The M-F1 score for the hatefulness model is around 50, whereas for offensiveness, it is around 80 (Barbieri et al. 2020). An additional factor is the TVBARC model itself, which may fail to converge if the data points are too large in magnitude. From above, one can observe that increasing the minimum offensiveness threshold does not greatly alter the structure of the time series, so it can be used to cut down on the scale as necessary (note that these threshold parameters can be freely tuned to a desired sensitivity). However, making the threshold too high may result in many 0 s in the series, which is especially true when filtering on hatefulness.

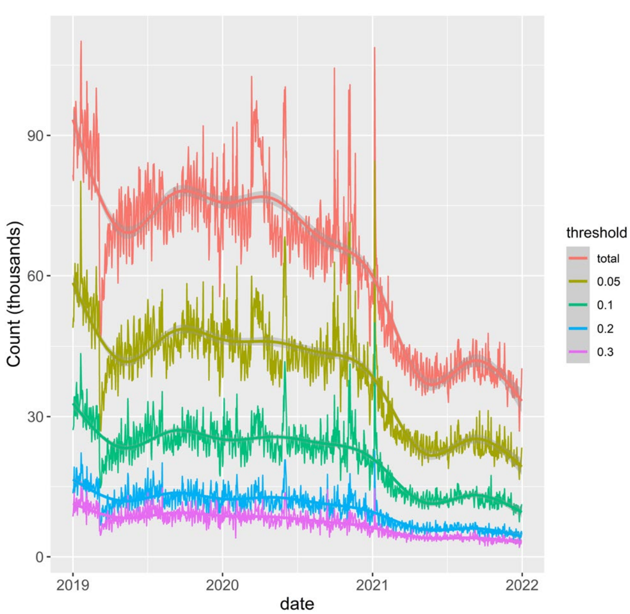

Fig. 4 Results of subsetting method using hatefulness

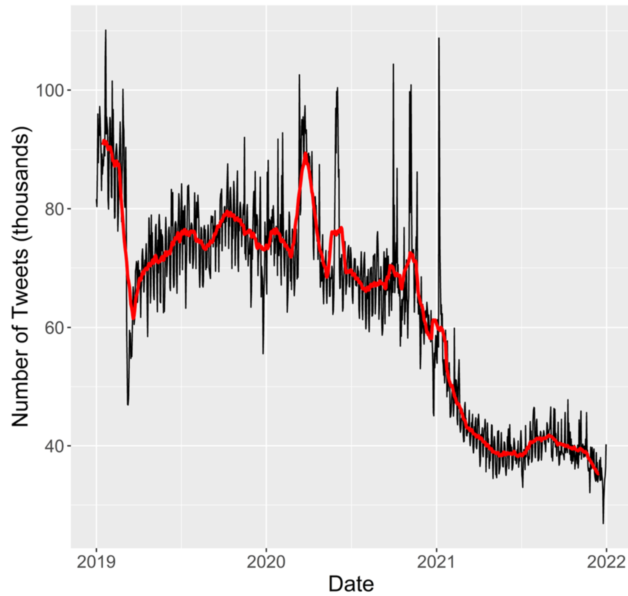

Fig. 2 Daily count of total tweets containing queried keywords, 2019–2021, superimposed by a 30-day centered rolling average

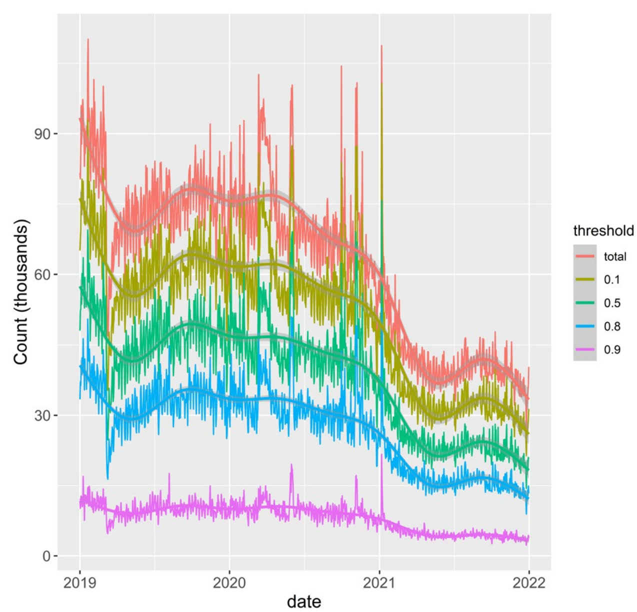

Fig. 3 Results of subsetting method using offensiveness

Not all offensive speech is considered cyberbullying, but hate speech can contain offensive speech like slurs, in this case directed to women or immigrants, which may be associated with cyberbullying. That being said, the data set has a nonzero number of tweets that lack offensive content but have a high probability of being hateful. Therefore, it would be wise to increase the threshold on offensiveness as necessary to allow model convergence and primarily focus on altering the hatefulness threshold.

In this study, we will use two separate thresholds for comparison. First, some notation is established. An x/y filter will refer to a filter that only accepts tweets with offensiveness probability greater than $\frac{x}{100}$ and hatefulness probability greater than $\frac{y}{100}$. In this notation, the two filters used are 25/0 and 25/50. Using these, of interest is rudimentary comparison with prior results before we fit to the TVBARC model. To motivate the use of this model, a visual analysis on the counts is performed.

4 Visual analysis on the raw and filtered counts

A visual analysis, similar to that done by Karmakar and Das (2020), allows one to deduce some trends and patterns. However, certain issues will limit the efficacy of such an analysis and justify implementation of a statistical model. First, we focus on the counts provided by the count endpoint with no thresholding, observed in Fig. 2.

Starting in 2019, one sees a sudden drop in counts, which levels out until the end of 2020. In 2020, there is a prominent peak in April, which roughly agrees with the findings of Karmakar and Das (2020). Throughout the rest of 2020, there are several spikes, but they are not persistent. When 2021 begins, the counts drop yet again, but unlike 2019, the counts stay at these lower levels.

In both 2019 and 2021, one observes a decrease in counts in the beginning of the year, though this does not occur in 2020. Instead, there are many large peaks though with an overall downward trend. In 2021, the downtrend accelerates and soon levels off. It is possible that with the advent of COVID-19, the typical downward trend was disrupted as lockdowns and social isolations precipitated increased internet use (Candela et al. 2020). The significant decrease may also be related to the findings of Bacher-Hicks et al. (2022) in their study, which uses Google search frequencies and bullying surveys to study the change in cyberbullying-related searches over time. The study shows that the log search intensity of school bullying and cyberbullying significantly decrease near the end of 2020 and beginning of 2021. But since one time series involves profanity and potentially hateful language on social media, and the other search engine data about bullying, it is possible that this is a coincidence. Additionally, due to the lack of thresholding, one cannot say that the frequency of cyberbullying events also follows the same trend.

We revisit the effects of thresholding, but now in the context of identifying trends. In Fig. 3, as the offensiveness threshold increases, the peaks shrink, and the trend begins to flatten. Note that it takes a threshold of 0.9 to induce some significant flattening. However, when thresholding on hatefulness, it only takes a threshold of 0.2 to achieve a similar degree of flattening. By the time we increase it to 0.5, the time series may appear constant, as shown in Fig. 5, which shows the different threshold settings in 2020. In both the raw and 25/0 time series, the peaks are easily identifiable, and thresholding on offensiveness works to reduce the scale while preserving the peaks. When using a large hatefulness threshold such as 0.5, prominent peaks disappear.

Fig. 5 Total counts and filtered counts in 2020

We are not necessarily satisfied with the analysis of the raw or 25/0 data since it may not represent the actual frequency of cyberbullying events. Further, it is also clear that applying any reasonable threshold on hatefulness may cause the time series to flatten out, making it difficult to identify trends visually. Therefore, to identify trends in abusive tweets, we must employ a statistical model.

5 Statistical modeling and analysis

While there are many models we can choose from to study this data, we opt to employ a time-varying Bayesian autoregressive count (TVBARC) model based on Poisson random variables (Roy and Karmakar 2020). There are many advantages to using this form of model, such as mitigating small sample-size, accounting for dependence, modeling subtle rather than abrupt change, and only needing a single parameter. More information on the motivations of the model, including in the context of Twitter data, is available in previous works of one of the authors (Karmakar and Das 2020; Das et al. 2020; Karmakar and Das 2021).

Suppose Z_t represents the true counts of all tweets, satisfying the query, that are as hateful and offensive as our threshold dictates. Let the time series provided by the count endpoint be Y_t. Now let p_t be the proportion of sample tweets on day t that meet the threshold requirements, with $0 \le p_t \le 1$ for all t. Then, we estimate series Z_t by

$$\hat{Z}_t = X_t = p_t Y_t$$

For this model, the conditional distribution for a count time series X_t given $\mathcal{F}_{t-1} = \{X_i : i \le (t-1)\}$ is

$$X_t \mid \mathcal{F}_{t-1} \sim \text{Poisson}(\lambda_t) \tag{1}$$

where

$$\lambda_t = \mu(t/T) + \sum_{i=1}^{p} a_i(t/T) X_{t-i} \tag{2}$$

Furthermore, $\mu(t/T)$ is the mean trend at time t, and $a_i(t/T)$ is the effect of the ith lag at time t. For the parameters that very over the time, a constraint on the parameter space is given as follows

$$\mathcal{P}_1 = \{\mu, a_i : \mu(x) > 0, 0 \le a_i(x) \le 1, \sup_x \sum_k a_k(x) < 1\}. \tag{3}$$

In order to sample for our desired parameters, we construct a likelihood function corresponding to (2) given by

$$L_1 \propto \exp\Bigg(\sum_{t=p}^{T} \Big[-\{\mu(t/T) + \sum_{i=1}^{p} a_i(t/T)X_{t-i}\} + X_t \log\{\mu(t/T) + \sum_{i=1}^{p} a_i(t/T)X_{t-i}\}\Big] - \sum_{j=1}^{K_1} \beta_j^2/(2c_2) - \sum_{l=0}^{p} \delta_l^2/(2c_1)\Bigg) \mathbf{1}_{0 \le \theta_{ij} \le 1} \tag{4}$$

The priors are given as

$$\mu(x) = \sum_{j=1}^{K_1} \exp(\beta_j) B_j(x), \tag{5}$$

$$a_i(x) = \sum_{j=1}^{K_2} \theta_{ij} M_i B_j(x), 0 \le \theta_{ij} \le 1, \tag{6}$$

$$M_i = \frac{\exp(\delta_i)}{\sum_{k=0}^{p} \exp(\delta_k)}, \quad i = 1, \ldots, p, \tag{7}$$

$$\delta_l \sim N(0, c_1), \text{ for } 0 \le l \le p, \tag{8}$$

$$\beta_j \sim N(0, c_2) \text{ for } 1 \le j \le K_1, \tag{9}$$

$$\theta_{ij} \sim U(0,1) \text{ for } 1 \le i \le p, 1 \le j \le K_2. \tag{10}$$

The B_j terms above are B-spline basis functions and the induced prior is supported by $\mathcal{P}$. Additional details on the restrictions, as well as a verification of the support, can be found in Karmakar and Das (2020), as the very same model is used. We apply this model to the estimation X_t of the true counts Z_t.

The fitting of the model over a count time series provides the mean trend of the series over time, as well as the coefficient values of the different autoregressive terms. From this one can deduce how the frequency of offensive/hateful tweets changed over time, as well as which lags are most related to each other. In the fitting procedure, we generate two models that differ in up to how many lags are represented. One contains up to lag 10, and the other 15. Generally, the mean trend captured is the same between these two models, though the width of the credible intervals may change. Exceptions to this will be mentioned when relevant.

5.1 Analysis of model fits on Twitter data

We now discuss the results of the model fits for each base using different threshold or lag settings. Figures for model fits are included in the appendix due to size. The top half of a given figure shows the value of $\mu(\cdot)$, given by Eq. (5). The grey bands for this function are its associated 95% credible intervals. The bottom half shows, for each lag i, the value of $a_i(\cdot)$ as in Eq. (6). When an individual lag i is mentioned in the discussion, it is referring to the value of $a_i(\cdot)$.

Beginning with 2019, there is no significant difference between the results of the lag 10 and lag 15 model, so we only focus on lag 10. Figure 8 displays the mean trend and coefficient values of the count data in 2019 by filtering on tweets with offensiveness greater than or equal to 0.25. As shown in Fig. 3, for low thresholds, this filter reduces the scale of the original data set, preserving all structure, so this can be interpreted as a scaled down version of the original count data. Here the credible intervals are quite small, so if the true counts were to be distributed by a 10-lag process, the mean trend would look like this. However, we also see no dependence on any lag besides the first. In the analogous model fit for the 25/50 data in Fig. 9, which focuses more on hateful tweets, we see a similar mean trend, capturing the peak between July and October. However, lag 7 becomes a lot more significant, and an earlier peak in the month is revealed. This process shows that the modeling procedure can uncover local extrema even when the process looks almost constant, like in Fig. 5 with the 2020 data.

The fits for 2020 are in Figs. 10 and 11. The trend captured between the two threshold settings is roughly the same. However, the credible intervals for the model passing only more hateful tweets are much wider. In the 25/50 setting, lags around 7 start to get a bit more significant. One may also directly compare this to previous results. In Karmakar's previous work, the peak of cyberbullying discourse occurred in April-May 2020, while in this setting, the peak in offensive and hateful tweets occurred later in the year, between July and October (Karmakar and Das 2020). A smaller peak near the beginning of the year is also observed. Compared to the previous study, the fits here have much wider credible intervals as well.

In the year of 2021, fitting the data through the 25/50 filter to the TVBARC model resulted in very different outcomes when using lag 10 (Fig. 12) as opposed to 15 (Fig. 13). For the 25/0 data, there is no major difference besides credible intervals and mean trend magnitude. There is an exceedingly rough similarity between the lag 10 and lag 15 model, but it is obfuscated by the very wide credible intervals in the lag 10 model. Yet in both models, lag 7 is important, as the coefficient value for this lag overtakes lag 1 during certain periods. In the lag 10 model, lag 6 is relevant, while in the lag 15 model, lag 13 is important. The results are similar, since if there is a weekly effect, one could also observe a "biweekly" relationship as well. Any dependence on lag p may cause lag $p + 7$ to be significant.

In general, when modeling the data on the 25/0, a scaled version of the raw counts, we most often do not see any weekly effect. However, when only considering tweets with hatefulness probability greater than 0.50, there is a very noticeable weekly effect, bringing out lags like as 5, 6, 7, 13, and 14. From this, we can infer that the frequency hateful tweets similar to those captured in the study have noticeable weekly seasonality that goes otherwise unobserved when looking at the raw counts. Additionally, the yearly pattern of two peaks in the beginning and end of the year roughly holds for 2019, 2020, and 2021. This implies that this time series also has a strong yearly seasonality. Furthermore, the mean of the hateful tweets, in most cases, follows a similar mean trend as the 25/0 series. Since this is just a scaled down version of the total data, the trend of hateful tweets mirrors that of the raw counts. This makes some sense since the filtering process takes a percentage of these counts. However, it may also reveal that the proportion of hateful tweets out of all tweets on a certain day remains about constant throughout time.

Next, correspondences with COVID-19 case counts are discussed in order to gauge possible relationships between the two time series.

5.2 Analysis in the context of the COVID-19 pandemic

By analyzing only tweets sourced from the USA, we may focus on COVID-19 counts from the same country. Depicted in Fig. 7 are the daily confirmed COVID-19 cases, taking a 7-day rolling average, provided by Ritchie et al. (2020). Beginning in 2020, there are consecutively larger peaks at around March, July, and October-December. We see a steep drop beginning near 2021, picking up during August-October. There is an exponential increase during the beginning of 2022.

As the visual analysis of the tweets in the study by Karmakar and Das (2020) did not necessarily reveal all information, it is advisable to fit the COVID data into some model to better compare to the modeled Twitter data. Fortunately, since daily new cases of COVID-19 are a count time series, we can use the same model as before to reveal information about the mean trend and significant lags. Using the same data set from Ritchie et al. (2020), we examine daily new cases in the US each day from January 23, 2020, to December 31, 2021 (since case data was not available early in the month of January 2020). We separately model the years of 2020 and 2021 and this time only use up to 10 lags. Note that because COVID-19 did not gain traction in the US in 2019, the modeled Twitter data in that year can be seen as a sort of experimental control for what the time series typically looked like pre-pandemic.

Knowing the pandemic began at around 2020, we compare each year's model to see if there are any changes. Starting in 2019 with Fig. 9, there is a peak in the first few months of the year, and then another during July-October. This same structure is also observed in 2020, as shown in Figs. 10 and 11, as well as 2021, in Figs. 12 (roughly) and 13. While the trend structure is similar, indicating the importance of seasonality, one distinction is that the secondary peak is largest in magnitude in 2020 (mean trend of 3000), followed by 2019 (about 2500) and then 2021 (about 1600). The largest peak occurring in 2020 is slightly corroborated by previous findings (Karmakar and Das 2020). Further, with each year the second peak persists for longer. This may be related to a variety of things, such as increased transmission during the winter, as COVID-19 proliferation rates can change significantly with temperature (McClymont and Wenbiao 2021), prompting more isolation and thus internet use, or increased cyber-aggression as a psychological response to the external stress of the pandemic (Wang et al. 2022).

Furthermore, a direct comparison between the mean trends can be made with each year of data to see if there are possible relationships. Consider the mean trend of tweets in Figs. 11 and 14. There are peaks in the mean trend at around October 2020, and the lag 1 coefficient follows a very similar pattern, oscillating with a peak in the summer and a

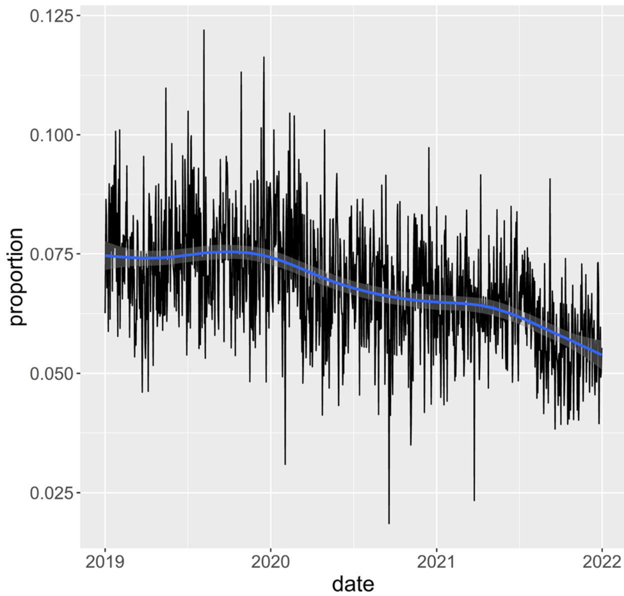

Fig. 6 Proportion of tweets meeting the 25/50 threshold out of the entire daily sample superimposed by a GAM smoother

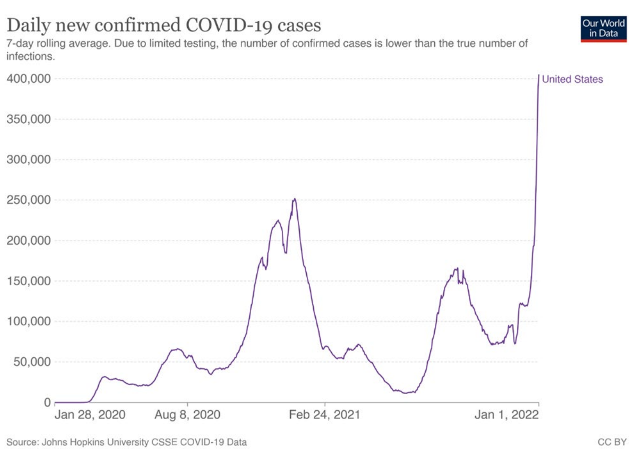

Fig. 7 7-day average of confirmed COVID-19 cases in 2020–2021

decrease in the fall. This offers some more credibility into the notion that potential cyberbullying events and COVID-19 cases increased in parallel. Now, for the year 2021 as shown in Figs. 12 and 15, one can observe the wider credible interval for each of the mean trends. There is a rough correspondence in peak trend in July, but it is obfuscated by the very wide intervals. More noticeable is the fact that both datasets are most influenced by lag 1 and 7, with the Twitter data set also containing other somewhat significant lags.

For the Twitter data, the mean trend structure's similarity across these three years begs the question of whether earlier results were more of a coincidence. More specifically, previous results show that cyberbullying discourse increased roughly the same time as COVID-19 cases began to rise (Karmakar and Das 2020). While these peaks occur at about the same time every year, it is important to note the change in magnitude throughout the years, dipping significantly in 2021. One may argue that the beginning of the pandemic may have brought about a sudden increase in abusive content, but as time progressed, the overall effect was to reduce such content. This angle is supported by findings using Google Trends (Bacher-Hicks et al. 2022).

It is also possible to study the proportions of hateful tweets, represented by p_t in the model definition. These proportions give a better idea as to the relative frequency of abusive events, as the previous time series studied can be affected by the overall number of active users. Figure 6 displays the proportion of daily tweets in the daily sample that meet the 25/50 threshold setting with a GAM smoother. The data are extremely noisy, but the smoother reveals a slight downward trend beginning just before 2020, suggesting a steady decrease in the proportion hateful content as the pandemic progressed.

6 Discussion

In this work, we use the official Twitter API's search endpoint to collect tweets contained keywords associated with cyberbullying over a period of 3 years. After running these tweets on an NLP model, we calculate the proportion of relevant tweets out of all tweets sampled that day using two metrics, hatefulness and offensiveness. It is our assumption that these metrics are useful in seeing whether a tweet could be considered cyberbullying or not, and that increased hateful content correlates with increased cyberbullying. The aforementioned proportion of tweets is then multiplied to the count endpoint's associated number, which is shown to be more reliable and thorough. From this we can construct a time series that is insensitive to the irregularities in daily sample sizes and better approximates the true number of hateful or offensive tweets.

The results show that the mean trend pattern of hateful tweets remains very similar through 2019, 2020, and 2021. Previous work posits that increased cyberbullying discourse happens in parallel with larger case counts (Karmakar and Das 2020), which can be seen in our own results with the parallel increase in mean trend in Figs. 11 and 14. However, we also notice that these peaks in the Twitter data occur in similar time frames in each of these years, including 2019 when the pandemic was not in full force. Other work suggests that cyberbullying may have been disrupted by COVID-19, as the log search intensity on Google of cyberbullying and bullying terms decreases in 2021 (Bacher-Hicks et al. 2022). Our work shows that the findings of Karmakar and Das (2020) may have been a result of lack of data and

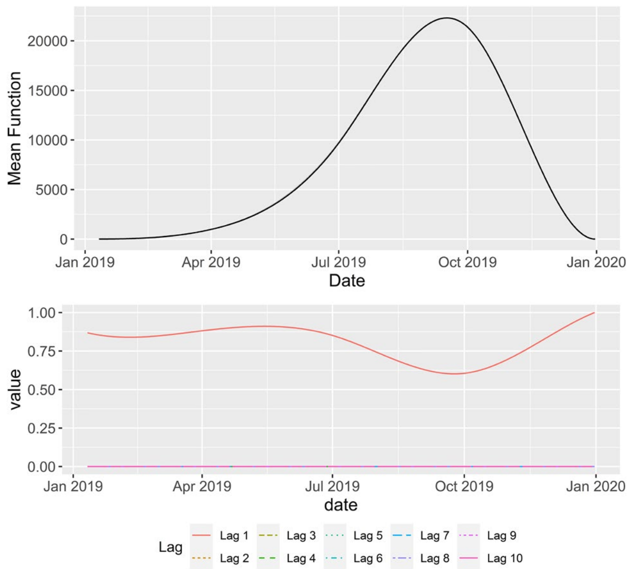

Fig. 8 Mean trend and AR coefficients of Lag 10 model on 2019 Twitter data with 25/0 filter

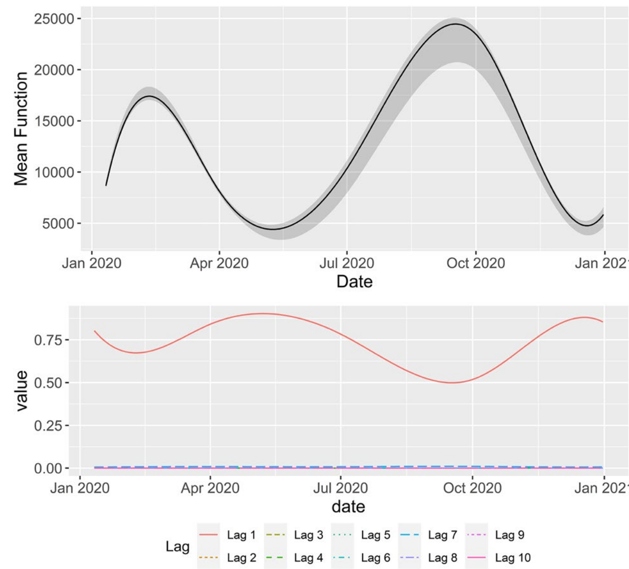

Fig. 10 Mean trend and AR coefficients of Lag 10 model on 2020 Twitter data with 25/0 filter

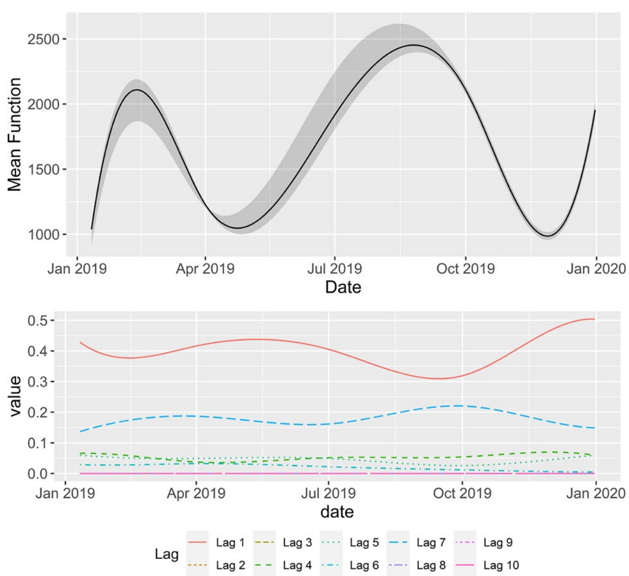

Fig. 9 Mean trend and AR coefficients of Lag 10 model on 2019 Twitter data with 25/50 filter

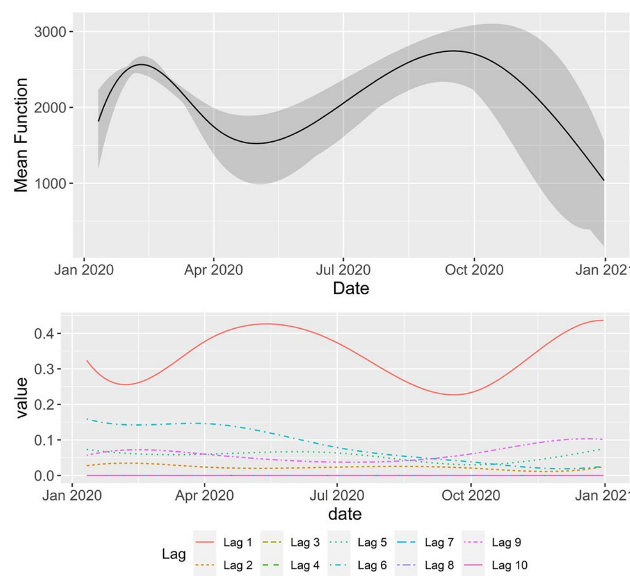

Fig. 11 Mean trend and AR coefficients of Lag 10 model on 2020 Twitter data with 25/50 filter

suggests that hateful content may have decreased overall, agreeing with Bacher-Hicks et al. (2022). One must take these correspondences with a grain of salt, due to the important differences between social media and search engine data (Li et al. 2021).

While the mean trend may have remained roughly similar, there are slight differences across the years. The secondary peak during the latter half of the year grows increasingly wide every year, which may be a consequence of virus proliferation in winter months and lifting of non-pharmaceutical interventions like lockdowns and mask mandates (McClymont and Wenbiao 2021; Singh et al. 2021). However, these findings may be a consequence of more active Twitter users, rather than a true increase in abusive content.

Additionally, from the important seasonal lags revealed in the model fit's AR coefficients, the seasonal effect is likely the most significant factor in determining the quantity of potential cyberbullying events. Further, both the trend of

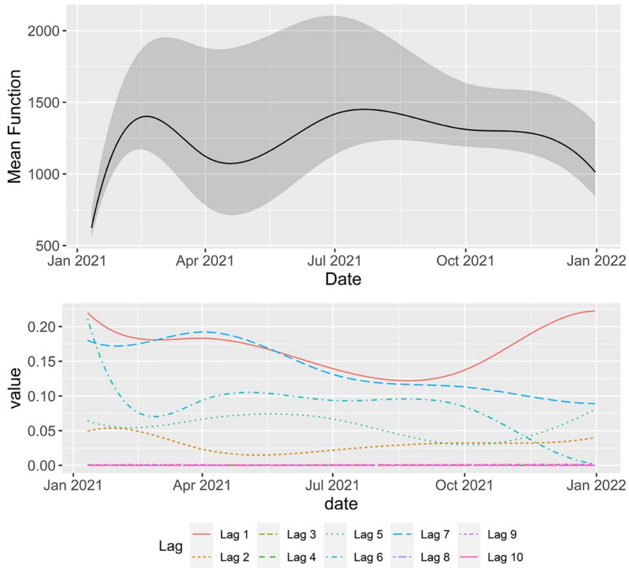

Fig. 12 Mean trend and AR coefficients of Lag 10 model on 2021 Twitter data with 25/50 filter

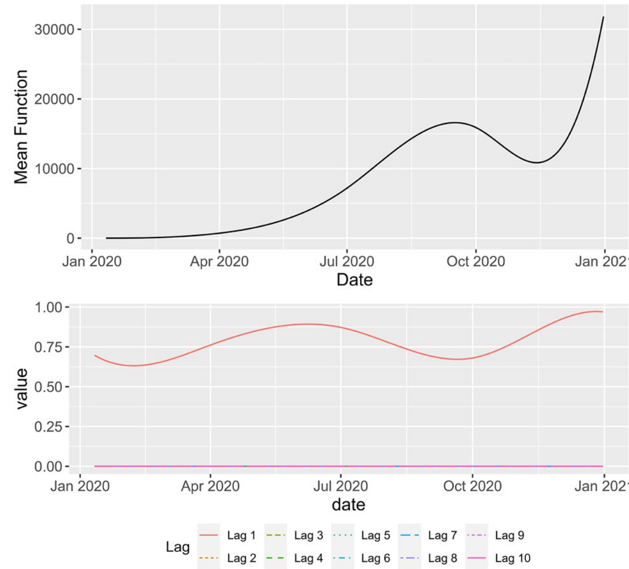

Fig. 14 Mean trend and AR coefficients of Lag 10 model on 2020 COVID-19 new case data

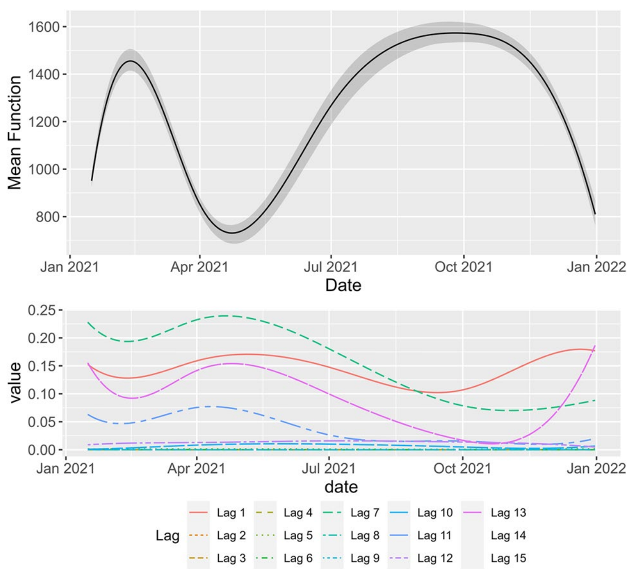

Fig. 13 Mean trend and AR coefficients of Lag 15 model on 2021 Twitter data with 25/50 filter

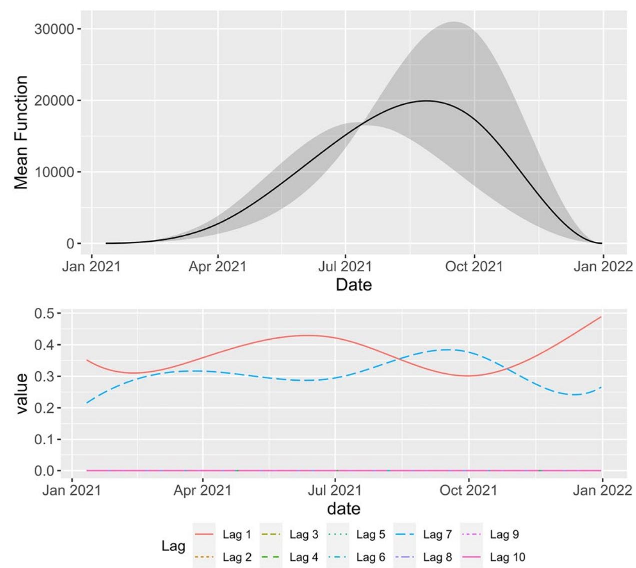

Fig. 15 Mean trend and AR coefficients of Lag 10 model on 2021 COVID-19 new case data

Twitter data and daily case counts in 2021 (Figs. 12, 15), are both heavily influenced by lags 1 and 7.

To summarize the possible effect of the pandemic, it is observed that the mean trend of tweets suddenly grows with the introduction of the pandemic but then cuts down significantly a year into it. Along this line, there is an increase in volume of potential cyberbullying tweets from 2019 to 2020, while there is a decrease from 2020 to 2021. This can be attributed to the raw number of hateful tweets decreasing as shown in Fig. 2, while the proportion of hateful tweets out of a given day's sample also decreases across the same time frame (Fig. 6). Both the Twitter and COVID-19 case time series are influenced by seasonal lags, indicating a weekly effect.

6.1 Limitations and suggestions for future research

A major limitation of this study is the uncertainty of how much hateful and offensive content correlates with cyberbullying events. This problem is exacerbated by racial biases in many training data sets online, the effects of which being noticeable in our own data. Further, it must be stressed that data of this nature in general cannot imply causal relationships between COVID-19 case counts and cyberbullying trends. As an example, note the finding of decreased hate speech quantity as well as proportion over the course of recent years. COVID-19 cases may not be a direct cause of this, and one may have to address several sources of confounding, such as increase internet usage precipitated by lockdown measures (Candela et al. 2020).

The study's immediate results can only be used to infer about the approximate state of cyberbullying events in the USA among English-speaking Twitter users. One also cannot be certain that the selection of keywords produces a good sample to pull out potential cyberbullying events from. While pulling geo-tagged tweets helps in collecting a more thorough sample (Morstatter et al. 2013), Twitter's sampling algorithm remains unknown and produces unique issues for time-series analysis, such as whether it maintains a fixed sampling rate to maintain structural relationships in the count series. Additionally, the NLP model used (Barbieri et al. 2020) is not necessarily state-of-the-art and was employed for its availability and ample documentation, rather than seeking the best performing algorithm known, incorporating more complex information such as social media structure and user information (Cheng et al. 2019a; Dadvar et al. 2013). Thus, the model's predictions on what it considers offensive or hateful may not be the most accurate. Further, the hate speech the NLP model is concerned about is primarily against women and immigrants as opposed to a broader scope of hate speech including, for instance, racism and homophobia. And again, the assumption that hateful tweets correlate with cyberbullying is not thoroughly justified.

Additionally, due to the importance of the weekly and yearly seasonality in the studied time series, employing a seasonal model in future work would better reflect the dynamics of the data. One may also consider integrating spatial methods as well, where the data set is broken down into several regions where the analysis is performed independently, like the work done by Babvey et al. (2021). This spatial analysis can be augmented with a similar daily count time series for a continuous analysis across several regions. It is also possible to use methods of Vector Auto-regression (VAR) to model several time series simultaneously, such as COVID-19 cases, abusive tweets, and search engine data. While not specifically studying cyberbullying, studies using VAR methods show the possibility of predicting suicides using search engine data (Taira et al. 2021) and COVID-19 cases with a great variety of variables (Wang et al. 2021b). A potentially challenging but rewarding avenue would be being able to maintain the ability to construct a count time series while employing more complex prediction models such as those in Cheng et al. (2019a); Dadvar et al. (2013). One may also consider a similar Bayesian analysis as in this paper while recruiting more time series variables from Ritchie et al. (2020), such as vaccinations, ICU admissions, and deaths, all of which are count data.

Appendix: Bayesian model fit graphs

See Figs. 8, 9, 10, 11, 12, 13, 14 and 15.

Acknowledgements We thank the Twitter team for granting us an academic research license to carry out this study.

Author contributions CP collected the data, performed the analysis, and wrote the manuscript text. SK provided code for the Bayesian model and offered guidance on manuscript structure and communication of results. All authors reviewed the manuscript.

Funding The second author's research is supported by NSF DMS 2124222.

Data availability The search endpoint data with modeled probabilities, count endpoint data, and the query used for the API requests are made available at Harvard Dataverse (https://dataverse.harvard.edu/privateurl.xhtml?token=8b866f9e-3a0c-4897-a839-b0fe6a0c2fc8). The code for the TVBARC model is available on GitHub (https://github.com/royarkaprava/TVBARC).

Declarations

Conflict of interest The authors report no conflict of interest.

References

Aboujaoude E, Savage Matthew W, Starcevic Vladan, Salame Wael O (2015) Cyberbullying: review of an old problem gone viral. J Adolesc Health 57(1):10–18

Babvey P, Capela F, Cappa C, Lipizzi C, Petrowski N, Ramirez-Marquez J (2021) Using social media data for assessing children's exposure to violence during the covid-19 pandemic. Child Abuse Neglect 116:104747

Bacher-Hicks A, Goodman J, Green JG, Holt M (2022) The covid-19 pandemic disrupted both school bullying and cyberbullying. Technical report, National Bureau of Economic Research

Barbieri F, Camacho-Collados J, Espinosa-Anke L, Neves L, (2020) TweetEval: Unified benchmark and comparative evaluation for tweet classification. In: Proceedings of Findings of EMNLP

Barlett Christopher P (2017) From theory to practice: cyberbullying theory and its application to intervention. Comput Human Behav 72:269–275

Barlett CP, Rinker A, Roth B (2021) Cyberbullying perpetration in the covid-19 era: an application of general strain theory. J Soc Psychol 161(4):466–476

Barlett CP, Simmers MM, Roth B, Gentile D (2021) Comparing cyberbullying prevalence and process before and during the covid-19 pandemic. J Soc Psychol 161(4):408–418

Basile V, Bosco C, Fersini E, Nozza D, Patti V, Pardo FMR, Rosso P, Sanguinetti M, et al (2019) Semeval-2019 task 5: multilingual detection of hate speech against immigrants and women in twitter. In: 13th International workshop on semantic evaluation, pp 54–63. Association for Computational Linguistics

Belchior Mota Daniela Cristina, Yury Vasconcellos, da Silva Thaís, Costa Aparecida Ferreira, Helena Magna, da Cunha Aguiar, Maria Eduarda de Melo Marques, and Ricardo Manes Monaquezi, (2021) Mental health and internet use by university students: coping strategies in the context of covid-19. Ciência & Saúde Coletiva 26:2159–2170

Bonanno Rina A, Shelley H (2013) Cyber bullying and internalizing difficulties: above and beyond the impact of traditional forms of bullying. J Youth Adolesc 42(5):685–697

Bozkurt A, Jung I, Xiao J, Vladimirschi V, Schuwer R, Egorov G, Lambert S, Al-Freih M, Pete J, Olcott Jr D et al (2020) A global outlook to the interruption of education due to covid-19 pandemic: navigating in a time of uncertainty and crisis. Asian J Distance Edu 15(1):1–126

Candela M, Luconi V, Vecchio A (2020) Impact of the covid-19 pandemic on the internet latency: a large-scale study. Comput Netw 182:107495

Cheng L, Li J, Silva YN, Hall DL, Liu H (2019) Xbully: Cyberbullying detection within a multi-modal context. In: Proceedings of the twelfth acm international conference on web search and data mining, pp 339–347

Cheng L, Guo R, Silva Y, Hall D, Liu H (2019) Hierarchical attention networks for cyberbullying detection on the instagram social network. In: Proceedings of the 2019 SIAM international conference on data mining, p 235–243. SIAM

Cornell D, Klein J, Konold T, Huang F (2012) Effects of validity screening items on adolescent survey data. Psychol Assess 24(1):21

Cortis K, Handschuh S (2015) Analysis of cyberbullying tweets in trending world events. In: Proceedings of the 15th International conference on knowledge technologies and data-driven business, pp 1–8

Dadvar M, Trieschnigg D, Ordelman R, De Jong F (2013) Improving cyberbullying detection with user context. In: Advances in information retrieval: 35th European conference on IR research, ECIR 2013, Moscow, Russia, 24–27 March, 2013. Proceedings 35

Das S, Kim A, Karmakar S, (2020) Change-point analysis of cyberbullying-related twitter discussions during covid-19. arXiv preprint arXiv:2008.13613

Davidson T, Bhattacharya D, Weber I (2019) Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516

Gayo-Avello D, Metaxas P, Mustafaraj E (2011) Limits of electoral predictions using twitter. In: Proceedings of the International AAAI conference on web and social media vol. 5, pp 490–493

Huang Q,Singh VK, Atrey PK (2014) Cyber bullying detection using social and textual analysis. In: Proceedings of the 3rd International workshop on socially-aware multimedia, pp 3–6

Jain O, Gupta M, Satam S, Panda S (2020) Has the covid-19 pandemic affected the susceptibility to cyberbullying in india? Comput Human Behav Rep 2:100029

Karmakar S, Das S (2020) Evaluating the impact of covid-19 on cyberbullying through bayesian trend analysis. In: Proceedings of the European interdisciplinary cybersecurity conference, pp 1–6

Karmakar S, Das S, (2021) Understanding the rise of twitter-based cyberbullying due to covid-19 through comprehensive statistical evaluation. In: Proceedings of the 54th Hawaii international conference on system sciences

Kowalski Robin M, Giumetti Gary W, Schroeder Amber N, Lattanner Micah R (2014) Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. Psychol Bull 140(4):1073

Kwan I, Dickson K, Richardson M, MacDowall W, Burchett H, Stansfield C, Brunton G, Sutcliffe K, Thomas J, (2020) Cyberbullying and children and young people's mental health: a systematic map of systematic reviews. Cyberpsychol Behav Soc Netw 23(2):72–82

Li Y, Goodell JW, Shen D (2021) Comparing search-engine and social-media attentions in finance research: evidence from cryptocurrencies. Int Rev Econ Financ 75:723–746

McClymont H, Wenbiao H (2021) Weather variability and covid-19 transmission: a review of recent research. Int J Environ Res Public Health 18(2):396

McHugh Meaghan C, Saperstein Sandra L, Gold Robert S (2019) Omg u# cyberbully! an exploration of public discourse about cyberbullying on twitter. Health Edu Behav 46(1):97–105

Mike T (2015) Evaluating the comprehensiveness of twitter search api results: a four step method. Cybermetr Int J Scientometr Informetr Bibliometr 18–19:1

Morstatter F, Pfeffer J, Liu H, Carley K (2013) Is the sample good enough? Comparing data from twitter's streaming api with twitter's firehose. In: Proceedings of the international AAAI conference on web and social media vol., 7, pp 400–408

Nand P, Perera R, Kasture A (2016) "how bullying is this message?": a psychometric thermometer for bullying. In: Proceedings of COLING 2016, the 26th International conference on computational linguistics: technical papers, pp 695–706

Olweus D, Limber Susan P (2018) Some problems with cyberbullying research. Curr Opin Psychol 19:139–143

Ritchie H, Mathieu E, Rodés-Guirao L, Appel C, Giattino C, Ortiz-Ospina E, Hasell J, Macdonald B, Beltekian D, Roser M (2020) Coronavirus pandemic (covid-19). Our World in Data. https://ourworldindata.org/coronavirus

Roy A, Karmakar S (2020) Bayesian semiparametric time varying model for count data to study the spread of the covid-19 cases. arXiv preprint arXiv:2004.02281, 19: 21

Signorini A, Segre A, Polgreen PM (2011) The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. PloS one 6(5):e19467

Singh S, Shaikh M, Hauck K, Miraldo M, (2021) Impacts of introducing and lifting nonpharmaceutical interventions on covid-19 daily growth rate and compliance in the united states. In: Proceedings of the National academy of sciences

Smith PK, Mahdavi J, Carvalho M, Fisher S, Russell S, Tippett N (2008) Cyberbullying: its nature and impact in secondary school pupils. J Child Psychol Psychiatry 49(4):376–385

Taira K, Hosokawa R, Itatani T, Fujita S, et al (2021) Predicting the number of suicides in Japan using internet search queries: vector autoregression time series model. JMIR Public Health Surveill 7(12):e34016

Talevi D, Socci V, Carai M, Carnaghi G, Faleri S, Trebbi E, di Bernardo A, Capelli F, Pacitti Francesca (2020) Mental health outcomes of the covid-19 pandemic. Rivista di psichiatria 55(3):137–144

Tumasjan A, Sprenger T, Sandner P, Welpe I (2010) Predicting elections with twitter: what 140 characters reveal about political sentiment. In: Proceedings of the international AAAI conference on web and social media

Wang Q, Luo X, Ruilin T, Xiao T, Wei H (2022) Covid-19 information overload and cyber aggression during the pandemic lockdown: the mediating role of depression/anxiety and the moderating role of confucian responsibility thinking. Int J Environ Res Public Health 19(3):1540

Wang Q, Zhou Y, Chen X, (2021) A vector autoregression prediction model for covid-19 outbreak. arXiv preprint arXiv:2102.04843,

Wiegand M, Ruppenhofer J, Schmidt A, Greenberg C (2019) Inducing a lexicon of abusive words–a feature-based approach. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, June 1-June 6, 2018, New Orleans, Louisiana, Volume 1 (Long Papers)

Yunhe W, Shi L, Que J, Lu, Qingdong Liu Lin, Lu Zhengan Xu, Yingying Liu Jiajia, Sun Y, Meng S et al (2021) The impact of quarantine on mental health status among general population in china during the covid-19 pandemic. Mol Psychiatry 26(9):4813–4822

Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R (2019) Semeval-2019 task 6: identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983