Using Machine Learning to Enhance Archival Processing of Social Media Archives

LIZHOU FAN, University of Michigan

ZHANYUAN YIN, The University of Chicago

HUIZI YU, Brown University

ANNE J. GILLILAND, University of California, Los Angeles

This paper reports on a study using machine learning to identify incidences and shifting dynamics of hate speech in social media archives. To better cope with the archival processing need for such large scale and fast evolving archives, we propose the Data-driven and Circulating Archival Processing (DCAP) method. As a proof-of-concept, our study focuses on an English language Twitter archive relating to COVID-19: tweets were repeatedly scraped between February and June 2020, ingested and aggregated within the COVID-19 Hate Speech Twitter Archive (CHSTA) and analyzed for hate speech using the Generative Adversarial Network (GAN)-inspired DCAP Method. Outcomes suggest that it is possible to use machine learning and data analytics to surface and substantiate trends from CHSTA and similar social media archives that could provide immediately useful knowledge for crisis response, in controversial situations, or for public policy development, as well as for subsequent historical analysis. The approach shows potential for integrating multiple aspects of the archival workflow, and supporting automatic iterative redescription and reappraisal activities in ways that make them more accountable and more rapidly responsive to changing societal interests and unfolding developments.

Additional Key Words and Phrases: Archival Processing, COVID-19, Generative Adversarial Network, Hate Speech, Machine Learning

1 INTRODUCTION

Cultural heritage informatics applies "a comprehensive, cross-disciplinary approach to supporting the entire lifecycle of cultural information and documentation procedures for the benefit of the preservation, study, and promotion of cultural heritage" [4, 57]. Open Archival Information System (OAIS)-based systems [47] and various interoperability standards [59] provide infrastructural specifications and promote such life cycle-based approaches. However, implementations remain underdeveloped in archival practice. The emerging transdisciplinary area of Computational Archival Science (CAS) seeks to change this situation by combining advanced computational and reasoning methods with archival thinking and testbeds in order to develop new tools and processing and analytical methods for archival administration [39]. Computational tools and methods that not only automate but also enhance multiple archival processes potentially will be particularly valuable in this respect.

To manage increasingly high volumes of born-digital and digitized content, computational tools and methods need to take into account the inter-dependencies and temporalities of different archival processes across the archival administrative workflow as well as the changing nature and value of open accession, dynamically evolving archival collections. Such applications must also satisfy overarching archival imperatives to ensure transparency, support accountability, and meet legal admissibility requirements for materials they preserve.

Beyond the above widely recognized requirements, we argue that computational tools and methods should provide new ways for archival practice to respond to two important movements in the archival field. One of these is its critical turn that is increasingly focused on surfacing and addressing issues of bias, misinformation, inclusion, racial equity, power and surveillance in archival administration, as well as in the content of archival

Authors' addresses: Lizhou Fan, lizhouf@umich.edu, University of Michigan, 105 S State St, Ann Arbor, MI, 48109; Zhanyuan Yin, The University of Chicago, 5801 S Ellis Ave, Chicago, IL, 60637, zyyin@uchicago.edu; Huizi Yu, Brown University, 69 Brown Street, Providence, RI, 02912, huizi_yu@brown.edu; Anne J. Gilliland, University of California, Los Angeles, 212 GSEIS Building, Los Angeles, CA, 90095, gilliland@gseis.ucla.edu.

collections [11, 12, 50]. Machine learning can be used to surface latent patterns of bias and underrepresentation in archival content, especially very high volume content that could not be analyzed manually. At the same time, however, if not carefully designed, unintended algorithmic biases or systematic biases based on prior knowledge may be introduced into archival administration and access processes [3, 5, 9, 55, 60]. In moving towards more implementations of socio-technical systems, such care is particularly important given that it is widely acknowledged that even non-automated archival appraisal and description practices inevitably introduce biases into the historical record and its representation [6, 46, 55]. The second movement is the growing recognition of the central roles that archives, as records and documentation generated in the course of bureaucratic and personal activities, can and do play in addressing societal grand challenges [2, 25], which is reflected in the increasing engagement of archival science with other emerging informatics areas such as crisis informatics [37, 38, 51] and human security informatics [27] in the development of socially and bureaucratically responsive and technologically flexible digital tools.

This paper presents proof-of-concept research that is situated at the nexus of these various developments. It is being carried out by a transdisciplinary team of academic researchers from computer and information sciences, statistics, digital humanities and archival studies. Our research directly addresses the questions raised in #1-3 of our larger research agenda below, as well as preliminarily contemplates others raised by #4-7 [69]:

- (1) *Knowledge discovery in newly archived digital content* – The utility of archived collections related to current affairs is often limited by how long it takes for the responsible archives, or archival steward to process the collections and make them available for research use. How can machine learning be used by archivists to surface latent knowledge as quickly as possible for use during an ongoing crisis, in controversial situations, or in decision or policy-making activities?
- (2) *Archival description* – How can machine learning processes assist archivists in identifying and describing latent knowledge contained in archived collections [21]? As new content is added to the collection, can iterative application of those processes update, recontextualize and generate increasingly granular descriptions of collection, making them continuously responsive to societal change?
- (3) *Archival appraisal* – Applied iteratively over time to open accession digital collections, could machine learning processes developed to enhance archival description also be applied to newly created digital content and thus support decision-making about whether or not to ingest it into the archives? Could these processes be developed further as tools for subsequent reappraisal of archived digital collections?
- (4) *Archival transparency* – How can algorithmic reasoning be made transparent to archivists and users of the archives?
- (5) *Archival bias* – How can the reasoning of machine learning processes be assessed by humans, when should such assessment occur, and how can that reasoning be corrected if it is found to be developing inappropriate biases?
- (6) *Hybrid archival processes* – What is the optimal balance between human processing (i.e., by archivists) and computational processing in the archival administration of high-volume digital collections? How might that balance shift over time as a result of the implementation of machine learning?
- (7) *End user tool development* – In what ways might computational tools designed to support archival practice also support scholarly and other user interpretation of archival content?

Specifically, this paper reports on a single study within this research: our attempts to identify incidence and shifting dynamics of hate speech from five months of English-language tweets relating to the COVID-19 pandemic that were made between February and June 2020. The tweets were repeatedly scraped, ingested and aggregated within the COVID-19 Hate Speech Twitter Archive (CHSTA) – an open accession collection (i.e., it will continue to acquire content from the same source) [23]. Our research suggests that it is possible to use machine learning and data analytics to surface and substantiate trends from CHSTA and similar social media

archives that could provide immediately useful knowledge for crisis response, in controversial situations, or for public policy development, as well as for subsequent historical analysis.

More broadly speaking, this study offers proof-of-concept of an approach that can assist archivists in comprehending the content and value of very high-volume digital collections such as those generated by social media or electronic mail, thereby supporting nuanced, or even corrective or reparative, redescription activities in ways that are not possible to do manually because of the scale of the collection in question. Moreover, archivists can factor the new and increasingly precise insights these methods yield into ongoing appraisal processes whereby newly created tweets are selected for archival ingestion into CHSTA, and, if later desired, into the reappraisal of previously archived tweets. Thus, the approach shows potential not only for integrating multiple aspects of the archival workflow, but also for supporting automatic iterative redescription and reappraisal activities in ways that can update existing collection practices and metadata and make them more responsive to changing societal interests and values, and unfolding developments and knowledge.

2 BACKGROUND

Because of the transdisciplinarity of our study, we refer here to a range of activities, including general archival appraisal practices, social media and COVID-19 specific archival research, and machine learning methods.

2.1 Automating and Iterating Archival Description and Enhancing Archival (Re)Appraisal

Regular knowledge-based archival techniques depend on the archivist's understanding of the medium, arrangement and structure of the records, as well as knowledge about the record creators and major topics discussed [68]. Archival description typically takes the form of multi-purpose representations of archival content (a.k.a. finding aids) that facilitate user access, as well as archival administrative control over collections [29]. Description first represents the entire collection and its provenancial and documentary contexts, and then successively details as many levels within a collection hierarchy as are deemed useful for access and also cost-efficient for archivists to produce. Item- or within-item description is uncommon. In other words, archives generally try to find a "sweet spot" between the archival labor involved in developing a detailed finding aid, and the anticipated usefulness of very granular description. Although digital encoding and dissemination of description and database-driven archival information systems are supported in many countries today by standards such as XML-compliant Encoded Archival Description (EAD) [20], description itself remains almost entirely a manual, value-added archival process.

A major limitation of such approaches to description, therefore, as well as the appraisal that determines which collections are accessioned or ingested, is the amount of human effort and inevitable bias involved in manually discerning noteworthy content in what are often very large and heterogeneous collections of material. Large backlogs and resource shortages have led archives to adopt descriptive practices that are faster to produce manually, but that contain even less detail [32]. Subsequently redescrbing or enhancing description based upon hindsight or new information, or for reparative purposes, while desirable, is often impractical to undertake manually. With very large-scale digital collections, particularly those generated by social media and other forms of digital communication, it becomes less and less possible to develop useful granular or indeed summary descriptions of content, or representations that are responsive to the shifting contexts within which that content is created and subsequently accessed and interpreted. Making content usable and actionable as soon as possible after its creation, e.g., for policy analysis, crisis remediation, or in response to a legal action, rather than several years later as is the current norm for archival practice, adds further pressure on manual archival processing. Finally, with demands for transparency and accountability in archival processes, archives are increasingly being called upon to explain how description and indeed appraisal decisions were arrived at [62, 66].

While some degree of human input into description is likely to remain desirable [22, 41], particularly for providing historical and cultural context, the overall descriptive effort could be ameliorated if certain aspects of those collections could be described automatically and if additional information about the collection as a whole that is not readily discernible to the human eye could be identified.

A number of approaches have been attempted working with official records generated through bureaucratic activities. As early as 1995, Wallace recognized the potential of using structural and procedural metadata inherent in the content itself as a mechanism whereby born-digital content, especially electronic records, could automatically, albeit partially, describe itself [63]. One early computational attempt was Collection-based Persistent Digital Archives [44]. Although this approach was primarily designed for digital preservation, its use of high performance computing to analyze the structure of items within a predominantly uniform archival collection and derive an XML DTD allowed anomalous items to be identified for additional human descriptive attention. This work was further developed using data grid technology and an Integrated Rule-Oriented Data System (iRODS) and a Storage Resource Broker that incorporated policy and information management information, including descriptive and other representation information. Nevertheless, the researchers acknowledged that assigning appropriate descriptive metadata through a knowledge procedure within a policy-based system is a labor-intensive task and likely non-scalable [45]. They proposed defining the features that are present within collection items and then using automatic feature extraction to index those items [45]. Combinations of a rule-based structural analysis and pattern matching have also been used in the archival appraisal of electronic mail [26].

With higher capacity of computing power and the advancement of algorithms, automated methods for capturing or deriving descriptive information have gradually become feasible and meaningful. Moreover, much of manual description involves locating, constructing, figuring, and re-figuring context, which is potentially infinite. Many kinds of context, however, are understandable, identifiable and derivable by machines in ways that are considerably beyond human capacities [19]. To identify and categorize context, archivists can take advantage of automated classification tools. Categorization methods are primarily based on two approaches – pattern matching and rule-based – which cluster documents according to word patterns and apply user-defined rules separately [40, 61]. Many specific automated disposition algorithms are used in both categorization methods, including text Bayesian analysis models such as Adaptive Probabilistic Concept Modeling (APCM) and Probabilistic Latent Semantic Analysis (PLSA) [54, 61]. Using test collections and making multiple passes, the models learn the document patterns and make classification decisions.

2.2 Archiving Social Media

Digital content generated through social media, although it may have certain predictable structural characteristics, is likely to be much more heterogeneous in its content than are bureaucratic records that are all generated out of the same organizational activity. Nevertheless, while it is not subject to the same policy rules and evidentiary scrutiny as bureaucratic records, the trustworthiness and transparency of archival administration of social media content may be a matter of public, and sometimes political concern, especially when it is used as a major organ of government communication [7]. Although many archives and libraries preserve social media content [22], there is a pressing need for more research and testing of computational descriptive methods that can not only eliminate some manual descriptive processes, but also surface new knowledge and increase contextualization. As early as 1992, a statistical content analysis demonstrated that a statistical approach could assist in archival description of the content of early social media implementations, and especially in surfacing latent patterns in that content [28]. The 2018 failure of the Library of Congress to be able to provide access to the entirety of its massive Twitter Archive resulted in a number of more selective efforts to archive and provide access to Twitter content [8]. In their examination of using platform APIs to identify social media content for preservation, Acker

and Kriesberg, however, cautioned that this approach presents a new evidentiary and administrative challenge for archivists because it can result in decontextualizing social media data streams [1].

2.3 COVID-19 Twitter Archives

In late January 2020, archivists and scholars around the world started collecting COVID-19-related Twitter content for a range of purposes. While some archives sought to document the overall distribution of public opinion on Twitter, many collections had specific focuses. For example, some archives cover certain geographic areas, such as the COVID-19 in Washington, D.C. Twitter Archive [16]; while others are collecting on a specific aspect related to the COVID-19 pandemic. While these archives have different purposes and operate at different scales, two trends can be noticed: the evolving standards for appraising (i.e., evaluating and selecting) content, and collaborative approaches to collecting.

Prior to the outbreak of COVID-19, little was known about the virus and its spread. Thus those seeking to identify and ingest relevant material into their new archives had little prior knowledge that could help them with identifying keywords for querying Twitter or predicting how much computing power and storage space would be needed for the resulting collection. For example, one of the earliest archives, COVID-19-TweetIDs [13], was initiated on January 28, 2020 and has been continuously updating the keywords it is using to collect. As it evolved from its original 14 keywords to the current 76 keywords (last updated on July 2020), the archive's collecting scope became massively enlarged. For non-consecutively collected archives relating to COVID-19, we should therefore expect considerable differences and evolution in content based on the evolving range of keywords that have been used to collect their content.

Documenting the Now [18] is a multi-institutional initiative that is developing archiving tools and protocols and building community practices to support archivists, activists and researchers who are working with social media data. Its crowdsourced guide to coronavirus documentation efforts indicates that a number of archives are collecting Twitter content, each collecting different pieces of the documentary record [17]. Since the Twitter API limits scraping rates, different efforts during the same time period may yield totally different contents, even when they use similar query keywords. The distribution of complete tweet datasets is prohibited by Twitter's community rules, so some tweets may subsequently disappear, while similar tweets collected by other archives may be captured and documented. This crowd-sourced, heterogeneous archival collection of Twitter IDs plays an important role, therefore, in mitigating the information loss that might otherwise result from the limitations imposed by Twitter's API.

2.4 Machine Learning for Evolving Hate Speech Archives

Machine Learning offers powerful techniques for opinion mining in general and hate speech detection specifically. Existing hate speech detection techniques can be divided into two main sets of approaches: classical machine learning methods and deep learning methods. The former primarily uses classical machine learning techniques such as Logistic Regression, Support Vector Machine (SVM), and Naïve Bayes classifiers [10, 15, 33, 42, 70]. Either Lexicon-based features are derived from, or a Word to Vector process is implemented for each tweet. The features or n-grams are then fed into a machine learning model to determine whether or not the tweet can be classified as hate speech. Deep learning primarily uses Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) [24, 43, 49, 70]. Tweets may be transformed into vectors and fed into the network either in the form of a CNN (matrix) or an RNN (sequence). The network is then trained based on the labelled dataset.

SVM and Logistic Regression are the most widely used methods, but recently deep learning methods have become more popular for hate speech detection. However, regardless of which methods are employed, we can identify a few problems they have in common. First, most of the hate speech detection techniques focus on detecting hate speech about racism or sexism in quite closely delineated contexts [15, 36]. Although a few

research studies have implemented hate speech detection with respect to a specific one-time event, few have focused on an ongoing phenomenon such as the COVID-19 pandemic. Thus, we believed it was important to develop a detection method for hate speech and hate speech trends relating to an event that might last for several months or more. Moreover, most hate speech detection techniques developed to date have been trained and tested on only a few small datasets that have been archived by some official institutions. One reason for this is that currently, as already noted, most of the hate speech datasets are identified and labelled manually. Because of varying institutional and personal standards for classifying hate speech, each tweet has to be read by three or more individuals to determine whether it indeed can be labelled as hate speech or not [15]. Some tweets have to be abandoned as they are controversial and a human decision cannot be reached. All of these issues restrict the size of the dataset that can be created. Thus, existing studies predominantly have used datasets containing approximately 10,000 tweets [70], which is not sufficient for deep learning. Aiming to solve this problem, we have focused on developing an algorithm that could auto-label the tweets, training the model recursively so that a larger dataset (or collection) can be built over time.

3 THE COVID-19 HATE SPEECH TWITTER ARCHIVE (CHSTA)

3.1 Data Scraping

Since the first discovery of the COVID-19 coronavirus in early 2020, the infectious respiratory disease quickly spread across the world. Because it was first identified in Wuhan, Hubei, China, "China" and "coronavirus" consequently became associated words in the public's mind. As COVID-19 developed into a pandemic, Asians and people of Asian descent increasingly became targets of racism and hate speech [14, 31, 52, 58]. In some severe cases, xenophobic rhetoric has led to anti-Asian violence and assaults [58].

Previous research [65] has demonstrated that increasing hate speech on social media leads to more crimes against minorities, including violence, harassment and criminal damage. Even though people are physically distant, the information shared through social media streams, including hate speech and derogatory comments, remains influential [56, 67]. To better understand hate speech's potential impact on public opinions and behaviors, and to effectively document and predict the occurrence of hate speech, we have been identifying and capturing tweets related to coronavirus and China in an archived collection and then analyzing them with a specific focus on hate speech. During the critical period between coronavirus' emergence and ensuing outbreak (January 31, 2020 to April 7, 2020), we conducted six waves of data collection, each lasting approximately one week. We queried the Twitter API using the keyword "coronavirus+and+china" and collected 3,482,197 unique tweets. The aggregated data comprise the COVID-19 Hate Speech Twitter Archive (CHSTA), which we have made publicly available on Github in order to encourage its use by other researchers¹.

¹Access CHSTA: <https://github.com/lizhouf/CHSTA>

3.2 Dataset Description

Table 1. Details of Tweets Collection

Tweet Collection			
Wave	Start Date	End Date	Total Number of Tweets
1	2020-01-31	2020-02-07	445,893
2	2020-02-07	2020-02-17	297,654
3	2020-03-03	2020-03-12	422,146
4	2020-03-15	2020-03-22	1,031,969
5	2020-03-22	2020-03-29	673,625
6	2020-03-31	2020-04-07	610,910

Fig. 1. New COVID-19 Cases by Date

Table 1 shows the number of tweets in CHSTA that were collected through each wave. Although the coverage of each wave differs by a few days, we find that wave 4 (March 15th to March 22th) has significantly more tweets compared to other waves. Combining Table 1 with Figure 1, which plots the number of daily new COVID-19 cases reported by WHO, we observe that wave 4 corresponds to the initial explosion in cases [35] as more testing kits become available. Thus it is highly likely that the newly aroused concern about public health garnered much online discussion.

Aggregating the data from the six waves, as Table 2 shows, we present the descriptive statistics of six variables that we use in our later analysis. Due to the large right-skewness of the distribution of each variable, we report the

minimum, median, maximum and standard deviation. The large standard deviation of each variable indicates the high heterogeneity of users who published tweets related to "china" and "coronavirus" during our data collection period. Furthermore, by analyzing the favorite count and retweet count of each tweet, we can gauge each tweet's impact in the online Twitter community.

Table 2. Descriptive Statistics of Selected Variables (6 waves combined)

Field	Minimum	Median	Maximum	Standard Deviation
favorite_count	3	3	1998	65.44
retweet_count	0	676	224801	10652.65
user_favorite_count	0	14478	2734656	72800.71
user_follower_count	0	465	76474977	518544.18
user_friend_count	0	661	1703101	7723.6
user_listed_count	0	2	379082	2274.79

Fig. 2. Selected Top 100 Hashtags in Wave 1 and 6

The rich CHSTA also records the hashtags used in each tweet. In Figure 2, we show the top 100 selected hashtags in wave 1 and wave 6 (some hashtags such as "coronavirus" and "COVID19" are removed due to redundancy). We note that the most popular hashtags in wave 6 contain more locations (e.g. "united states", "italy", "india", "france", "iran") compared to the top hashtags in wave 1. Furthermore, by continuously scraping and populating CHSTA, we capture the change in overall sentiment towards Chinese people and the coronavirus. Hashtags such

as "chinavirus", "chinesevirus", and "wuhanvirus" are among the top hashtags in later waves. In Figure 3, we find a surge in the use of derogatory hashtags in wave 4 (March 15th to March 20th), accounting for a notable 10.88 % of the frequency of top 100 hashtags. Although some of the tweets that contain derogatory hashtags express disapproval of such language use, a substantial number of tweets employ these hashtags when expressing anti-Chinese sentiments on the Twitter platform. The continued, aggrandized circulation of rhetoric such as "chinese virus" distorts the origin and nature of the public health threat by wrongly implying that Chinese people and people of Chinese origin are the cause of the pandemic.

Fig. 3. Percentage of Racist Hashtags ("chinesevirus", "chinavirus", "wuhanvirus") Across Waves

4 THE DATA-DRIVEN AND CIRCULATING ARCHIVAL PROCESSING (DCAP) METHOD

In this section, we describe the DCAP method in detail, including intuition, key terminologies, structure, and the alternative lexicon base for appraisal (the Hatebase Dictionary).

4.1 Intuition of the DCAP Method

When archiving an unfolding event, it is appropriate to use iterative corpus curation processes, such as Richard Rogers' proposals of "keyword snowballing" and "query snowballing" [48, 53]. Reviewing CHSTA and examining other COVID-19 Twitter archives underscore that archival appraisal and description of the tremendous number of tweets that are being collected is currently very time-consuming. This is a fundamental problem for those who wish to use such kinds of collections to understand more about the dynamics of an unfolding crisis or other situation. We therefore propose the Data-driven and Circulating Archival Processing (DCAP) method as a possible way to address the research questions provided in the Introduction section, including those related to processing an evolving archive, providing immediate insights into crisis dynamics, supporting public policy and decision-making, and recording shifts in social perception to facilitate subsequent historical analysis.

The core idea of the DCAP method is inspired by Generative Adversarial Network (GAN). With a GAN, the goal is to build a generator to produce figures that can cheat the discriminator. In the model, a **generator**, **G**, and a **discriminator**, **D**, are trained iteratively. New figures are fed into the discriminator and trained to improve the

discriminator's accuracy, while the discriminator's response to the figures generated by the generator provides feedback to and optimizes the generator. The two processes above are executed repeatedly so that both the discriminator and the generator are iteratively improved [30]. Our method borrows this idea, but instead of building a generator that could produce new figures, we seek to build a generator that would produce a new dictionary that can better identify hate speech in relation to a specific topic. The rest of our algorithm, discussed below, follows the structure of the GAN network, but is revised in detail ².

Fig. 4. Workflow of the DCAP Method

As shown in Figure 4, there are five iterative steps:

- *Collecting Tweets via Twitter API*. Using either a fixed set of keywords or an evolving set, archivists can collect tweets around a certain topic. The collection criteria we used are discussed in Section 3.1.
- *Archival Appraisal (by User or by Algorithm)*. Based on archivists' domain knowledge or descriptions used for related archives, the tweets are initially labelled as either "hate speech" or "non-hate speech". Specifically, we use a crowdsourced Twitter dataset [64] to provide us with necessary prior knowledge.
- *Expanding the Hatebase Dictionary*. The original Hatebase dictionary is based on the Hatebase [34], which includes thousands of discriminatory words. From the tweets labelled as "hate speech" in each appraisal iteration, we extract useful words and use them to augment the original Hatebase dictionary. Archives addressing other topic areas could use representative dictionaries in those fields instead.

²The source code is publicly available: <https://github.com/lizhouf/DCAP>

- *Data-Driven Archival Description.* As discussed in Section 3.2, some data-driven archival descriptions can be created based on the archived data itself – for Twitter, many fields are included in metadata when scraping. With the updated Hatebase Dictionary, we are also able to visualize Hate Speech. From the perspective of archival description, this information is particularly useful to researchers who want to find different groups of Tweets based on certain conditions. And these conditions become easier to understand with visually presented data.
- *Creating or Improving the Selection Algorithm.* A selection algorithm is created in the first iteration and updated thereafter. We will discuss the algorithms we use in detail in Section 4.3.

4.2 Key Terminologies of the DCAP Method

Base Dataset and Training Dataset:

One main difference between the DCAP Method and GAN Network is that for the GAN Network, all of the images in the dataset are labelled. However, in our case, few tweets about COVID-19 have labels. Therefore, we need an initial dataset to train the initial discriminator.

We name our datasets as $T_0, T_1, T_2, \dots, T_n$, where T_0 stands for the base dataset that is labelled by T_0^L and $T_1, \dots, T_n$ stand for training datasets that are fed into the network in iterative steps. This dataset is used to train the initial discriminator to start our generative model.

Discriminators:

In the DCAP Method, discriminators are used to detect whether or not a tweet is controversial. We denote discriminators as $D_0, D_1, \dots, D_n$. D_0 is the base discriminator that is directly trained by the base dataset (T_0, T_0^L), and $D_1, \dots, D_n$ are datasets that are trained iteratively in the model.

Discriminators could be expressed in such general form:

$$\hat{T}^L = D(T) = \text{step} \left(\sum_{i=1}^k p_k d_k(T) \right) \quad (1)$$

$$\text{step}(x) = \begin{cases} 1 & x \geq 0.5 \\ 0 & x < 0.5 \end{cases} \quad (2)$$

here, D is a meta-classifier composed of k sub-classifiers, $d_1, \dots, d_k$. p_k stands for the weight of each sub-classifiers, determined by the precision score of each classifier:

$$p_k = \frac{TP}{TP + FP} \quad (3)$$

The reason we choose the precision score is that in our trials, we find that the classifiers tend to classify more non-controversial speech than controversial speech (False Positive). Thus, we would like to reinforce the model to put more attention on avoiding classifying non-controversial speech as controversial ones.

As for the sub-classifiers $d_1, \dots, d_k$, we use Logistic Regression (d_1), multi-SVM (d_2), ANN (d_3), Random Forest (d_4), and Gaussian Naïve Bayes (d_5) as sub-classifiers.

Generators:

In the DCAP Method, Generators are used to produce new Hatebase dictionaries that are in turn used to distinguish (label) controversial speech. We denote generators as $G_0, G_1, \dots, G_n$. G_0 is the base generator that is directly drawn from the base dataset as well as labels (T_0, T_0^L). The rest of the generators produce new dictionaries iteratively in the training process. We use the Naïve Bayes method to build new Hatebase dictionary based on the conditional probability of each word appearing in labelled controversial speech and labelled non-controversial speech. The detailed algorithm to produce a dictionary from a dataset as well as its labels, (T, T^L) is the following:

Algorithm 1 Algorithm of Generators

Require: Tweets T , Tweets' Labels T^L , Accept Threshold k , Original Hatebase Dictionary $dict_o$;

Ensure: New Hatebase Dictionary $dict_n$;

- 1: Pick all words that appeared in T and create word set W ;
- 2: Calculate the probability that controversial speech in (T, T^L) :

$$P_c = \frac{|T[i]^L == +|}{|T|}$$

- 3: Delete words appeared in 91,000 common word dictionary and get new word set W' ;
- 4: **for** $w \in W'$ **do**
- 5: Calculate the probability that the word appeared in labelled controversial speech:

$$P_{cond} = \frac{|(T[i]^L == +) \cap (w \in T[i])|}{|T[i]^L == +|}$$

- 6: Calculate the probability of the word appeared in all tweets:

$$P_{full} = \frac{|w \in T[i]|}{|T|}$$

- 7: The probability of w belonging to controversial speech is calculated by

$$P = \frac{P_c P_{cond}}{P_{full}}$$

- 8: **if** $w \in dict_o$ and $(|w \in T[i]| \leq 5$ or $P < kP_c)$ **then**
- 9: Add w to $dict_n$, $dict_n[w] = 0$;
- 10: **else**
- 11: **if** $|w \in T[i]| > 5$ and $P \geq kP_c$ **then**
- 12: Add w to $dict_n$, $dict_n[w] = P$; $=0$

To explain in detail, the algorithm of the generators is based on Naïve Bayes Theorem, i.e., we pick all the words that appeared in the tweets and exclude those belonging to the common words dictionary (reference). Then for the rest of the words, we calculate the probability of the word appearing in the controversial speech instance. Then we use the Naïve Bayes Theorem:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \quad (4)$$

Here, event A stands for the event "the tweet is controversial", and event B stands for the event "word w appeared in the tweet". Thus $P(A|B)$ stands for the probability that a tweet is controversial if a given word w appeared in it, which is what we desired. Furthermore, to raise our standard of creating the Hatebase dictionary to ensure that the size of that dictionary will not be overly large, we set two standards for adding the word to the dictionary:

- (1) The word has to appear more than 5 times in the tweets: $|w \in T[i]| \geq 5$, unless the word is in the original Hatebase dictionary. The threshold of 5 times is a predefined parameter that is used to facilitate machine understanding of important hate words while saving computing time.
- (2) The Bayes probability has to be greater than a threshold: $P(A|B) > kP(A)$. k is pre-determined and equal to 4 in our case. It means that the probability that the word appeared in the controversial tweets is k times greater than that of it appearing in the overall tweets.

Also, we create the affiliation score AF_w of each word w belonging to the controversial dictionary equal to $P(A|B)$. The higher the affiliation score, the higher the probability of the word appearing in an instance of controversial speech. Thus, the words associated with higher affiliation scores are better indicators of controversial speech.

4.3 Structure of DCAP Method

Having introduced the discriminator and generator, we need a structure to combine the two to create a generative flow such that we can iteratively train our network. The algorithm is as follows:

Algorithm 2 Algorithm of DCAP Method

Require: Tweets Dataset $T_0, T_1, \dots, T_n$; Original Hatebase Dictionary $dict_0$;

Ensure: Discriminators $D_0, D_1, \dots, D_n$; Generators $G_0, G_1, \dots, G_n$; Hatebase Dictionaries $dict_1, \dots, dict_{n+1}$;

- 1: Start Round 0;
 - 2: Use original Hatebase dictionary to count number of hate words in each tweet in the original dataset T_0 ;
 - 3: Train the base discriminator D_0 using original dataset T_0 (**number of hate words as one feature**);
 - 4: Renew the Hatebase dictionary using generator G_1 to get $dict_1$;
 - 5: **repeat**
 - 6: For Round i , add new dataset T_i to the network;
 - 7: Count number of hate words in each tweet for dataset $T_0, \dots, T_i$ using dictionary $dict_i$;
 - 8: Train new discriminator D_i using dataset $T_0, \dots, T_{i-1}$ (as T_i has no label);
 - 9: Label dataset T_i by discriminator D_i and concatenate datasets $T_0, \dots, T_i$;
 - 10: Use datasets $T_0, \dots, T_i$ to create new Hatebase dictionary $dict_{i+1}$;
 - 11: **until** All datasets are fed to the model =0
-

Because of the speciality of the initial dataset (the only dataset that is labelled using human labor), we separate the initialization step from the main body of the method. That is, we use the original dataset to build up a base discriminator D_0 such that we can improve its ability in further training steps. In this discriminator, we are more tolerant of the Type II Error than the Type I Error, as the ability of the discriminator is improved, so it will become more acceptable to miss some of the controversial tweets than to mis-classify some of the non-controversial tweets. Then, we could produce the first updated Hatebase dictionary $dict_1$ by integrating the classified result using G_0 . The rest of the model can be divided into two parts.

Part 1: $G \rightarrow D$

This step is used to produce new discriminators. Currently, assuming we have generator G_i and produce dictionary $dict_{i+1}$, we need to use the dictionary to update the datasets. The key point that connects the updating process of the dictionaries and the training process is that we are using the number of hate words as one feature

since it shows significance in training the model. This additional feature could improve the R^2 of the meta-classifier from 0.72 to 0.84 on average. Here, we define the number of hate words in one tweet as equal to the sum of affiliation scores of all words in the tweet if they appear in the Hatebase dictionary. Therefore, for tweet $T[j] \in T$:

$$N_j = \sum_{w \in \text{dict}_i \text{ and } w \in T[j]} AF_w \quad (5)$$

Then, we can train a new discriminator D_{i+1} based on the updated dataset $T_0, \dots, T_i$. Note that for the newly added dataset T_{i+1} , it is unlabelled (i.e. no y value), so it can not be used in this step of training. Instead, we pass the dataset to the trained discriminator D_{i+1} and label the data. Therefore, these datasets can be used to update the new dictionary.

Part 2: $D \rightarrow G$

This step is used to produce new Hatebase dictionaries. One assumption we made here is that we trust our discriminator that it labels the tweets correctly, especially the controversial tweets. Also, as the newly added dataset T_{i+1} is labelled by D_{i+1} , we can add the dataset to our original dataset $T_0, \dots, T_i$ to update the dictionary and get dict_{i+1} , which would be used in the next $D \rightarrow G$ process. The full generating process can be referred to the generators in the previous part.

4.4 Alternatives to the Hatebase Dictionary

Based on our trial, we find that the generator will usually produce hate words aggressively. That is, in a few cases, some words that are not actually offensive are included, based on Naïve Bayes Theorem. To avoid introducing bias to the algorithm, we decide not to impose any prior knowledge when distinguishing hate words. As a side effect, some non-controversial words are included as indicators of controversial speech if they appear significantly more times in those tweets that are labelled as controversial than those that are not. However, in accordance with the Law of Large Numbers, when the number of tweets in the dataset is sufficiently large:

$$P(P(\text{controversial} | w \in \text{tweet}) = \mathbb{E}(w \in \text{controversial tweet})) = 1 \quad (6)$$

where there is still probability that bias may occur. Therefore, we create two additional dictionaries to address this occurrence. In total, we have three dictionaries, including *MB* (Machine-defined Bigram), *MU* (Machine-defined Unigram), and *HU* (Human-selected Unigram). As shown in Figure 5, *MB* is separately generated, while *MU* is a subset of *HU*:

Fig. 5. Relationships of the Three Dictionaries

Specifically, *HU* is a subset of *MU* based on human annotation. Instead of labelling a large dataset of tweets, we can focus on cleaning the dictionary, which is only about 500 to 2000 words. We could either clean the dictionary after each round of training, or after the overall training process. Either way, retaining this human element, even if small, is nevertheless one check for algorithmic skew over time.

Taking a different approach, *MB* constructs bigram word pairs instead of a unigram. We scan through the tweets and calculate bigrams such that both words contained in the bigram appear in the Hatebase dictionary. The affiliation score of the bigram is the product of the affiliation scores of the two words. Although it may shrink the size of the dictionary, this is a more conservative way to expand the dictionary, since commonsense would suggest that it would be really rare that both words should not be in the Hatebase dictionary. Also, even if one of the words is irrelevant as a hate word, if the other one has an offensive meaning, it is still likely that the phrase constructed by putting the two words together is offensive.

4.5 Summary of the DCAP Method

Borrowing the idea from GAN and applying it to Natural Language Processing (NLP), the DCAP method undertakes iterative corpus refinement and brings machine learning into the mix for discovery, identification and understanding. In particular, our DCAP method has various advantages, especially when compared with traditional NLP methods:

- (1) The DCAP method allows researchers to train on partially labelled data. In NLP, data preprocessing and manually labeling data can be costly and may have low reliability. However, the DCAP method tries to decrease the need for human-labelled data to the lowest amount possible. In our model, only a few thousand human-labelled tweets are needed to start up its iterative process, and the large portion of unlabelled data, formerly unable to be used, can be labelled by our model and input during further training.
- (2) The DCAP method is an iterative method that can continually input data and train models for each iterated step. In practice, it is suitable for the data that is released or updated periodically: when the data is updated, it can be input to the model that has been trained using the previous data. As the model is easy to update, it can serve as a monitor to supervise trends in the data we are studying.
- (3) In addition to traditional NLP techniques such as sentiment analysis, DCAP also uses machine learning techniques that allow researchers to build a more specific model related to a topic. Only the first step of the DCAP model relies on the open-source hate speech dictionary, and the following iterative steps build the model's own hate speech dictionary into a new dictionary related to a more specific topic.
- (4) Since the DCAP method contains two iterative models training each other, the two models can also function separately. The discriminator can be used to predict whether a tweet belongs to controversial speech, while the generator can generate hate speech dictionaries based on labelled tweets.

5 RESULTS

Fig. 6. R^2 of testing data (black), ratio of *MB* to *MU* (red), and newly added "hate words" (histogram) over six rounds

Using the DCAP Method for distinguishing controversial speech about COVID-19 in CHSTA, we increasingly improve the accuracy of the model and obtain an improvement of hate speech detection rate from 89.92% to over 99%, as shown in Figure 6. Therefore, Our DCAP Method has demonstrated that it is capable of distinguishing hate speech relating to COVID-19 in a large open accession archival collection. This suggests the utility of similar data-driven archival processing for other topic areas also.

In previous work, many models have only focused on detecting hate speech, which means that it is necessary to develop labelled data to train the model. Those archived datasets are usually labelled manually, which means it is not feasible for large datasets. Our method uses iterative learning between two models to label the whole dataset step-by-step. To appraise and describe a large dataset on a topic about which there is little previous knowledge, we only need to manually label a small part of the data to start our DCAP Method, which realizes semi-supervised learning of hate speech detection.

Furthermore, our discriminator uses bagging (aggregating the result of multi-classifiers) to make predictions, which avoids the common error that current models usually make: focusing on accuracy rather than precision. Although some other current models have high accuracy, they tend to label more non-controversial speech as controversial, which may prove problematic for archivists, activists, and subsequent researchers, not to mention policy and decision-makers who may need to respond to a trend in the moment.

Our model also realizes a step-by-step improvement of both the discriminator and the generator (as well as the dictionary). One of the reasons why we use a generative model instead of a direct model that would feed all the datasets into the model is that our base discriminator and dictionary is trained based on a small dataset, and is yet to be comprehensive. Some obvious hate words or new kinds of hate words may not be included. Thus, if we use

iterative learning, the dictionary will be expanded step-by-step, and more hate words can be identified, especially as additional scraped data is added, which will improve the ability of the discriminator. The following table shows the number of new words added to the dictionary and the number of words in total after each iteration:

Number of <i>MU</i> in Original Dictionary				
474				
Round	Number of New <i>MU</i>	Number of Total <i>MU</i>	Number of Total <i>MB</i>	<i>MB/HU</i>
0	182	656	NA	NA
1	152	808	384	47.52%
2	74	882	467	52.95%
3	55	937	508	54.22%
4	92	1029	565	54.91%
5	52	1081	624	57.72%
6	42	1123	662	58.95%

Table 3. Change of *MU* and *MB* Across Iterations

Therefore, according to the idea of GAN, the discriminator and generator (dictionary) will improve each other during each round of training. Moreover, in the steps of generating the models, we also extend *MB*, *HU*, and *MU* lexicon resources for Hate Speech. For example, in the preparing iteration, the dictionaries are updated as shown in Appendices I and II. These dictionaries are useful for the research of domain knowledge and future constructions of archives with similar topics. Furthermore, Table 3 reveals the relationship between the numbers of *MU* and *MB* over iterations. If the ratio of *MU* to *MB* is calculated, we can observe that *MU/MB* is approximately between 0.5 to 0.6 and is increasing throughout the iterations. It reveals that when more *MU* are discovered, the unigram dictionary is expanded and more related controversial bigrams *HU* can potentially also be discovered. This indicates that the expansion of the unigram hate dictionary is also meaningful for the expansion of the *n*-gram hate dictionary for a larger *n*.

Another advantage of DCAP can be observed via checking the confusion matrix of the test set for each iteration. In Table 4, we can focus on the cells that are colored light blue. They represent the False Positive cases, which means the number of non-controversial tweets that are classified as controversial by the model. We can observe that throughout the iterations, the False Positive cases follow a declining trend. The precision of the model is defined as:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

and the precision of our aggregate model rises from 43.25% to 99.31%, which is a significant improvement. This result indicates that DCAP helps avoid mis-classifying some non-controversial speech instances. Thus, our discriminator is capable of picking out controversial speech without unduly influencing people's daily Twitter usage.

	Round 1			Round 2		
		Actual			Actual	
		C	N		C	N
Predicted	C	2888	3789	C	3472	600
	N	590	36179	N	502	48036
	Round 3			Round 4		
		Actual			Actual	
		C	N		C	N
Predicted	C	3755	255	C	4138	146
	N	262	49338	N	269	50057
	Round 5			Round 6		
		Actual			Actual	
		C	N		C	N
Predicted	C	4565	162	C	4934	34
	N	193	50690	N	376	51266

Table 4. Confusion Matrix after 6 Iterations (C stands for Controversial; N stands for Non-controversial)

6 CONCLUSION

When processing archives with characteristics such as large volume or dynamic updates, traditional analytical tools, such as basic statistical summaries and text processing visual help are no longer sufficiently informative or consistently accurate. With the advancement of algorithms and computational resources, the application of machine learning tools such as ensemble learning, deep neural networks, and reinforcement learning, becomes increasingly accessible and applicable. However, because of the complexity and heterogeneity of born-digital archives, and of social media archives in particular, direct applications of machine learning tools without fine modification is risky in terms of ensuring the trustworthiness of the archival processes. Inappropriate use of "black box" technology is likely to add to existing archival biases and reduce the transparency and potentially the usability of these archival collections. This is especially likely to be the case when taking into account specific attributes of social media archives such as the incompleteness of data entries, the randomness of data collection, and the involvement and possible dynamism of different languages and their grammars.

It is essential, therefore, as reflected in our model, to combine machine learning with NLP techniques. Our DCAP model borrows the idea of Generative Adversarial Network (GAN) in deep learning. The model has a discriminator and a generator, which is similar to the structure of GAN. In particular, the discriminator takes the role of generating predictions and the generator takes the role of producing new datasets to improve the functionality of the discriminators. Although we didn't change conventional machine learning techniques, such as the random forest in the discriminator training process, in the circulation process between the discriminator and generator, the instances are represented as tweets (information in natural language) instead of images in matrix form. In addition, adhering to the fact that the objects to be classified are tweets and the classification criteria are based on controversial words, the generator is adapted to monitor the appearance frequency of controversial words by producing a controversial speech dictionary such that the next training process of the discriminator can be better supervised through the dictionary. For its counterpart in the GAN where new data sets are generated and fed into the model, our model feeds the data into the model progressively in rounds and uses the discriminator of each round to label the newly added data. Besides taking advantage of GAN so that the training process is iterative and keeps improving, we also overcome the problem that only a small portion of the tweets are labelled manually: turning a supervised learning problem into a semi-supervised learning problem.

If we revisit the overall deployment of the DCAP method, we can clearly see its straightforward advantages in archival processing, especially for social media archives, while some highly rigid attributes of machine learning are carefully taken care of. On the one hand, the implementation of machine learning is essential, as it can efficiently handle large scale data archives. It can also generate rules and produce predictions from training data to testing data, or even unknown future data. On the other hand, when applying machine learning methods, the forms of input and output are restricted, and sometimes further requirements are posed for the dataset. To overcome the disadvantages, it is important to borrow from machine learning techniques in flexible ways and make modifications based on data structure, data type, purpose, theme, or usage. The key idea behind the combination of machine learning and processing this kind of digital archival collection is to follow the rationality of machine learning methods and create models with high specificity for individual data archives, which, of course, is applicable to the specific requirements of social media archives.

Overall, we believe that this study demonstrates the potential of augmenting current manual archival appraisal and description processes in the following ways: it can provide rapid insights into specified facets and trends of social media that simply could not be achieved through human processes because of the scale of the collection; it provides iterative updating of appraisal and description of open accession social media collections that is highly and rapidly responsive to dynamic content and user needs; it is able to tackle critical and time-sensitive topics such as increasing societal racism that can be difficult to discern and call out; and it can render its analyses as visualizations that can provide users with alternative representations of collection content to that of traditional textual finding aids. At the same time, it holds archiving institutions and the algorithms they are employing accountable for how they arrive at their decisions about what they will collect, and how they have chosen to describe their collections and why.

7 DISCUSSION AND FUTURE WORK

As discussed above, machine learning and modeling methods iteratively help with adding content to the overall description of the archive, while careful consideration of the accountability of the process is essential. With the limitations on sharing full Twitter data, there is a corresponding limited number of full tweet datasets available, while most publicly available Twitter datasets of hate speech have the majority of their tweets deleted and thus we are unable to hydrate the tweet contents. Such incomplete sharing leads to two main problems in our study. First, the seed round of training data has a limited precision. Although we find the dataset in Waseem and Hovy's paper is suitable to use as a proof-of-concept, this dataset has 16,914 tweets pre-labelled for hate speech with an F1 feedback score between 70% and 75% [64]. Second, the evaluation process is too specific to include other datasets, and thus, the performance can be mostly properly evaluated using the confusion matrix in Table 4.

From the perspective of testing the performance, it is tolerable to have these limitations, if we mainly regard our study as a proof-of-concept of performing the DCAP method and our research is carried out in a timely manner. However, in our future research, we want to address these limitations and to improve the current DCAP method. For our next step, we will enable a feedback loop so that we can better understand what we should collect over time and build in additional accountability for collecting decisions (i.e., improving archival appraisal). The machine learning and modeling process itself is actually a record of appraisal reasoning which takes into consideration the changing context of the archive over time as records aggregate and reflect an ever more complete but continually evolving documentary context. This appraisal process also addresses the issue of lack of evaluation methods and can be even further improved using hybrid human-computer supervision.

In this way, the result of the Machine Learning augmented appraisal process will not only enlarge the body of detected incidences of hate speech, but also increase the knowledge about the content and importance of the archive (i.e., the process also augment archival description), while at the same time producing an explanation of the reasoning behind how tweets were identified and classified as hate speech (or not). From the perspective

of understanding and contextualising what counts as hate, the use of machine learning can enable a broader spectrum of problematic content. Machine learning can then situate hateful content within this spectrum based on the updating prior knowledge, rather than just focusing on the most hateful contents. Such a rich and informed approach to archival processing may additionally shed light on optimal balances between human and automated labor. It will also point to possible end user tools that would further exploit metadata associated with the content of social media archives.

8 ACKNOWLEDGEMENT

The co-authors would like to thank the anonymous reviewers for their insightful comments and suggestions on this manuscript, as well as the UCLA Scholarly Innovation Lab for supporting data collection and data management.

REFERENCES

- [1] Amelia Acker and Adam Kreisberg. 2020. Social media data archives in an API-driven world. *Archival Science* 20, 2 (01 Jun 2020), 105–123. <https://doi.org/10.1007/s10502-019-09325-9>
- [2] Micah Altman and Chris Bourg. 2018. A Grand Challenges-Based Research Agenda for Scholarly Communication and Information Science. *MIT Grand Challenge PubPub Participation Platform* (17 Dec 2018). <https://doi.org/10.21428/62b3421f> <https://grandchallenges.pubpub.org/pub/final>.
- [3] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20, 3 (2018), 973–989. <https://doi.org/10.1177/1461444816676645> arXiv:<https://doi.org/10.1177/1461444816676645>
- [4] Association for Information Science and Technology (ASIS&T) [n.d.]. *Introducing Cultural Heritage Informatics into the Curriculum of LIS Education*. Retrieved 2020-10-26 from <https://www.asist.org/meetings-events/webinars/introducing-cultural-heritage-informatics-into-the-curriculum-of-lis-education/>
- [5] Solon Barocas and Andrew D. Selbst. 2016. Big Data’s Disparate Impact. *California Law Review* 104, 3 (2016), 671–732. <http://www.jstor.org/stable/24758720>
- [6] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity, Cambridge, United Kingdom.
- [7] Kathleen Margaret Brennan. 2010. Believe Me: Authenticity, Federal Social Media Use, and the Problematic Record in the American Digital Public Sphere. *Critical Library and Information Studies Special Issue* 2, 2 (2010).
- [8] Axel Bruns. 2018. The Library of Congress Twitter Archive: A Failure of Historic Proportions. <https://medium.com/dmrc-at-large/the-library-of-congress-twitter-archive-a-failure-of-historic-proportions-6dc1c3bc9e2c>
- [9] Taina Bucher. 2018. *IF... THEN: Algorithmic power and politics*. Oxford University Press, Oxford, United Kingdom.
- [10] Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data science* 5, 1 (2016), 11.
- [11] Michelle Caswell and Marika Cifor. 2019. Neither a beginning nor an end : Applying an ethics of care to digital archival collections.
- [12] Michelle Caswell, Ricardo Punzalan, and T-Kay Sangwand (Eds.). 2017. *Special issue on Journal of Critical Library and Information Studies* (Jul 2017).
- [13] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill* 6, 2 (29 May 2020), e19273. <https://doi.org/10.2196/19273>
- [14] Stephen M Croucher, Thao Nguyen, and Diyako Rahmani. 2020. Prejudice toward Asian Americans in the COVID-19 pandemic: the effects of social media use in the United States. *Frontiers in Communication* 5 (2020), 39.
- [15] Thomas Davidson, Dana Warmesley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *International AAAI Conference on Web and Social Media* (2017). <https://arxiv.org/pdf/1703.04009.pdf>
- [16] DC Oral History Collaborative [n.d.]. *COVID-19 in Washington, D.C. Twitter Archive*. Retrieved (10.27.2020) from <https://digdc.dclibrary.org/islandora/object/dcplislandora%3A237558>
- [17] DocNow [n.d.]. *Documenting COVID-19*. Retrieved (10.27.2020) from <https://docs.google.com/document/d/1v5ts08spFq6SpW53h2OJULcdRoPEbyl6xpah31kW-H0/edit>
- [18] DocNow [n.d.]. *Documenting the Now*. Retrieved (10.27.2020) from <https://www.docnow.io/>
- [19] Wendy M. Duff and Verne Harris. 2002. Stories and names: archival description as narrating records and constructing meanings. *Archival Science* 2, 3-4 (2002), 263–285.
- [20] Encoded Archival Description [n.d.]. *Encoded Archival Description*. Retrieved (10.27.2020) from <https://www.loc.gov/ead/>
- [21] Wolfgang Ernst and Jussi Parikka. 2013. *Digital memory and the archive*. University of Minnesota Press.
- [22] Lizhou Fan, Huizi Yu, and Anne J. Gilliland. 2021. #StopAsianHate: Archiving and Analyzing Twitter Discourse in the Wake of the 2021 Atlanta Spa Shootings. *Proceedings of the Association for Information Science and Technology* 58, 1 (2021), 440–444. <https://doi.org/10.1002/pra2.475>
- [23] Lizhou Fan, Huizi Yu, and Zhanyuan Yin. 2020. Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter. *Proceedings of the Association for Information Science and Technology* 57, 1 (2020), e313. <https://doi.org/10.1002/pra2.313>
- [24] Björn Gambäck and Utpal Kuma Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-Speech. *Proceedings of the First Workshop on Abusive Language Online* (2017), 85–90.
- [25] Anne J. Gilliland. 2015. Permeable Binaries, Societal Grand Challenges, and the Roles of the Twenty-first-century Archival and Recordkeeping Profession. *Archifacts* (Dec 2015), 12–30.
- [26] Anne J. Gilliland. 2016. Designing Expert Systems for Archival Evaluation and Processing of Computer Mediated Communications: Frameworks and Methods, Anne J. Gilliland, Sue McKemmish, and Andrew J Lau (Eds.). *Research in the Archival Multiverse (Melbourne: Monash University Press)*, 685–721.
- [27] Anne J. Gilliland and James Lowry. 2019. Human security informatics, global grand challenges and digital curation. *International Journal of Digital Curation* 14, 1 (Oct 2019). <https://doi.org/10.2218/ijdc.v14i1.636>

- [28] Anne Gilliland-Swetland and Carol Hughes. 1992. Enhancing Archival Description for Public Computer Conferences of Historical Value: An Exploratory Study. *The American Archivist* 55, 2 (1992), 316–330. <https://doi.org/10.17723/aarc.55.2.48530jm12848r7p2> arXiv:<https://doi.org/10.17723/aarc.55.2.48530jm12848r7p2>
- [29] Anne J. Gilliland-Swetland. 2001. Popularizing the Finding Aid. *Journal of Internet Cataloging* 4, 3-4 (2001), 199–225. https://doi.org/10.1300/J141v04n03_12 arXiv:https://doi.org/10.1300/J141v04n03_12
- [30] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. *the 27th International Conference on Neural Information Processing Systems 2* (2014), 2672–2680.
- [31] Angela R. Gover, Shannon B. Harper, and Lynn Langton. 2020. Anti-Asian Hate Crime During the COVID-19 Pandemic: Exploring the Reproduction of Inequality. *American journal of criminal justice : AJCJ* (07 Jul 2020), 1–21. <https://doi.org/10.1007/s12103-020-09545-1> PMC7364747[pmcid].
- [32] Mark Greene and Dennis Meissner. 2007. More Product, Less Process: Revamping Traditional Archival Processing. *The American Archivist* 68, 2 (Aug 2007), 208–263. <https://doi.org/10.17723/aarc.68.2.c741823776k65863> arXiv:https://meridian.allenpress.com/american-archivist/article-pdf/68/2/208/2052984/aarc_68_2_c741823776k65863.pdf
- [33] Edel Greevy and Alan F. Smeaton. 2004. Classifying Racist Texts Using A Support Vector Machine. *27th annual international ACM SIGIR conference on Research and development in information retrieval* (2004), 468–469.
- [34] Hatebase Inc. 2020. Hatebase. <https://hatebase.org/>, Last accessed on 2020-04-26.
- [35] Michelle A. Jorden et al. 2020. Evidence for Limited Early Spread of COVID-19 Within the United States, January–February 2020. *MMWR Morb Mortal Wkly Rep* 69 (2020), 680–684.
- [36] Irene Kwok and Yuzhou Wang. 2013. Locate the Hate: Detecting Tweets against Blacks. *Twenty-Seventh AAAI Conference on Artificial Intelligence* (2013).
- [37] Mark Latonero and Irina Shklovski. 2011. Emergency Management, Twitter, and Social Media Evangelism. *IJISCRAM* 3 (Jan 2011), 1–16. <https://doi.org/10.4018/jiscrm.2011100101>
- [38] Mark Latonero and Irina Shklovski. 2014. Crisis Informatics. *Faculty Publications* (2014). <https://doi.org/10.4172/2167-0587.1000e115>
- [39] Myeong Lee, Yuheng Zhang, Shiyun Chen, Edel Spencer, Jhon Dela Cruz, Hyeonggi Hong, and Richard Marciano. 2017. Heuristics for assessing Computational Archival Science (CAS) research: The case of the human face of big data project. In *2017 IEEE International Conference on Big Data (Big Data)*. 2262–2270. <https://doi.org/10.1109/BigData.2017.8258179>
- [40] R Kirk Lubbes. 2003. So you want to implement automatic categorization? *Information Management* 37, 2 (2003), 60.
- [41] Heather MacNeil. 1995. Metadata strategies and archival description: Comparing apples to oranges. *Archivaria* 39 (1995), 22–32.
- [42] Anu Masso and Tayfun Kasapoglu. 2020. Understanding power positions in a new digital landscape: perceptions of Syrian refugees and data experts on relocation algorithm. *Information, Communication & Society* 23, 8 (2020), 1203–1219. <https://doi.org/10.1080/1369118X.2020.1739731> arXiv:<https://doi.org/10.1080/1369118X.2020.1739731>
- [43] Yashar Mehdad and Joel Tetreault. 2016. Do Characters Abuse More Than Words? *SIGDIAL 2016 Conference* (2016), 299–303.
- [44] Reagan Moore, Chaitan Baru, Arcot Rajasekar, Bertram Ludaescher, Richard Marciano, Michael Wan, Wayne Schroeder, and Amarnath Gupta. 2000. Collection-Based Persistent Digital Archives - Part 1. *D-Lib Magazine* 6, 3 (Mar 2000). <https://doi.org/10.1045/march2000-moore-pt1>
- [45] Reagan Moore, Arcot Rajasekar, Michael Wan, Wayne Schroeder, Antoine de Torcy, Sheau-Yen Chen, Mike Conway, and Hao Xu. 2015. *Concepts in Distributed Data Management or History of the DICE Group*. Retrieved (10.27.2020) from <https://irods.org/uploads/2015/01/DICE-History.pdf>
- [46] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. <http://www.jstor.org/stable/j.ctt1pwt9w5>
- [47] OAIS Reference Model [n.d.]. *The fundamental standard for digital preservation*. Retrieved 2020-10-26 from <http://www.oais.info/>
- [48] Cathy O’Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Books, Largo, MD.
- [49] Ji Ho Park and Pascale Fung. 2017. One-step and Two-step Classification for Abusive Language Detection on Twitter. *First Workshop on Abusive Language Online* (2017), 41–45.
- [50] Holger Pötzsch. 2018. Archives and identity in the context of social media and algorithmic analytics: Towards an understanding of iArchive and predictive retention. *New Media & Society* 20, 9 (2018), 3304–3322. <https://doi.org/10.1177/1461444817748483> arXiv:<https://doi.org/10.1177/1461444817748483>
- [51] Christian Reuter and Marc-André Kaufhold. 2018. Fifteen years of social media in emergencies: A retrospective review and future directions for crisis Informatics. *Journal of Contingencies and Crisis Management* 26, 1 (2018), 41–57. <https://doi.org/10.1111/1468-5973.12196> arXiv:<https://doi.org/10.1111/1468-5973.12196>
- [52] Katherine J. Roberto, Andrew F. Johnson, and Beth M. Rauhaus. 2020. Stigmatization and prejudice during the COVID-19 pandemic. *Administrative Theory & Praxis* 42, 3 (2020), 364–378.
- [53] Richard Rogers. 2019. *Doing Digital Methods*. MIT Press, Cambridge, MA.
- [54] James Santangelo. 2009. RISE OF THE MACHINES. *Information Management* 43, 6 (2009), 22.

- [55] Nick Seaver. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4, 2 (2017), 2053951717738104. <https://doi.org/10.1177/2053951717738104> arXiv:<https://doi.org/10.1177/2053951717738104>
- [56] Tomer Simon, Avishay Goldberg, and Bruria Adini. 2015. Socializing in emergencies—A review of the use of social media in emergency situations. *International Journal of Information Management* 35, 5 (2015), 609–619.
- [57] Society of the American Archivists [n.d.]. *Cultural Heritage Informatics*. Retrieved (10.27.2020) from <https://www2.archivists.org/dae/kent-state-university/cultural-heritage-informatics>
- [58] Hannah Tessler, Meera Choi, and Grace Kao. 2020. The Anxiety of Being Asian American: Hate Crimes and Negative Biases During the COVID-19 Pandemic. *American Journal of Criminal Justice* (2020), 1–11.
- [59] The Digital Information LifeCycle Interoperability Standards Board [n.d.]. *DILCIS Board*. Retrieved 2020-10-26 from <https://dilcis.eu/>
- [60] Nicol Turner-Lee, Paul Resnick, and Genie Barton. 2019. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. <https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>
- [61] William Underwood and Richard Marciano. 2019. Computational Thinking in Archival Science Research and Education. In *2019 IEEE International Conference on Big Data (Big Data)*. 3146–3152. <https://doi.org/10.1109/BigData47090.2019.9005682>
- [62] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (Jun 2017), 76–99. <https://doi.org/10.1093/idpl/ix005> arXiv:<https://academic.oup.com/idpl/article-pdf/7/2/76/17932196/ix005.pdf>
- [63] David A. Wallace. 1995. Managing the Present: Metadata as Archival Description. *Archivaria* 39 (Feb. 1995), 11–21. <https://archivaria.ca/index.php/archivaria/article/view/12064>
- [64] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, 88–93. <http://www.aclweb.org/anthology/N16-2013>
- [65] Matthew L. Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2019. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology* 60, 1 (Jul 2019), 93–117. <https://doi.org/10.1093/bjc/azz049> arXiv:<https://academic.oup.com/bjc/article-pdf/60/1/93/31634412/azz049.pdf>
- [66] Stacy Wood, Kathy Carbone, Marika Cifor, Anne J. Gilliland, and Ricardo Punzalan. 2014. Mobilizing records: re-framing archival description to support human rights. *Archival Science* 14, 3 (01 Oct 2014), 397–419. <https://doi.org/10.1007/s10502-014-9233-1>
- [67] Bo Xie, Daqing He, Tim Mercer, Youfa Wang, Dan Wu, Kenneth R. Fleischmann, Yan Zhang, Linda H Yoder, Keri K Stephens, Michael Mackert, et al. 2020. Global health crises are also information crises: A call to action. *Journal of the Association for Information Science and Technology* (2020). <https://asistdl.onlinelibrary.wiley.com/doi/epdf/10.1002/asi.24357>
- [68] Geoffrey Yeo. 2013. Archival description in the era of digital abundance. *Comma* 2013, 2 (2013), 15–26.
- [69] Zhanyuan Yin, Lizhou Fan, Huizi Yu, and Anne J. Gilliland. 2020. Using a Three-step Social Media Similarity (TSMS) Mapping Method to Analyze Controversial Speech Relating to COVID-19 in Twitter Collections. In *2020 IEEE International Conference on Big Data (Big Data)*. 1949–1953. <https://doi.org/10.1109/BigData50022.2020.9377930>
- [70] Ziqi Zhang and Lei Luo. 2018. Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. *CoRR abs/1803.03662* (2018). arXiv:1803.03662 <http://arxiv.org/abs/1803.03662>

APPENDICES

Appendix I: Round 0 Machine-defined Unigrams & Human-selected Unigrams (Contained) (Some Hate Speech occurrences are marked red with parts masked for the sake of the scholarly audience)

Table 5. *MU* and *HU* (Round 0)

Machine-defined Unigrams					
Human-selected Unigrams					
hoe	murder	delusional	wilw	stealing	unite
shame	ruined	treated	horse	isis	yell
sick	brutal	feminist	wounded	timid	ads
fat	ill	prick	weight	degrading	im
crap	hell	shutuplayla	slapped	await	imma
richch**ga	freak	script	complain	deaths	mad
bigsh*txtalker	sh*t	override	sexism	treating	rape
kowtows	sickening	hospitals	veto	jealous	madness
rot	bastards	spreading	mmpadellan	suspected	trick
bum	nasty	sequel	quinncy	chaddi	hurt
coward	threats	normani	gang	presidential	shouting
fools	cruel	steal	iraqi	refused	harassment
bombs	terrifying	punched	weed	codes	murdered
hating	cowering	village	loveeeeeeee	illness	ialiyah
damned	fool	troops	cdnpoli	bluntofloud	trolling
ni**as	scumbag	pants	lil	requires	choke
annoying	bastard	disappointed	filters	shoots	shut
disgusting	anal	suspect	swaggmiiyaa	blames	terrorists
bi*ches	hated	rascals	terrorism	victim	lesbian
liars	horror	lose	thehill	fatal	
bad	retarded	pathological	hacked	slams	
idiotic	shaming	fear	urgent	senator	
idiots	rubbish	punch	treat	loyal	
fatuous	evil	blamed	j	htt	
dishonest	rewarded	sue	youngthug	mattmfm	
sl*t	as*hole	hate	pricks	troll	
cheated	devil	insult	consecutive	violent	
lied	f*ckers	screaming	piglets	exposing	
arse	damn	whine	normelanin	bombing	
heartless	sh*ts	prez	wolf	wars	
disgusted	pig	scream	impeach	worthless	
liar	retards	squldz	impeached	terrorist	
bomb		bloodbath	death	skittles	

Appendix II: Round 0 Machine-defined Bigrams (Some Hate Speech occurrences are marked red with parts masked for the sake of the scholarly audience)

Table 6. *MB* (Round 0)

Machine-defined Bigrams			
worthless idiot	evil bi*ch	shame hate	ni**as loveeeeeeee
murder victim	damn hell	sl*t shaming	evil heartless
guinea pig	bum bi*ches	disgusting pricks	heartless bastards
death threats	pathological liars	bi*ches steal	hate liars
pu*sy bad	damn codes	nasty pu*sy	damn heartless
bad bi*ch	sick sh*t	damn shame	idiot impeached
lose weight	bi*ches hate	nasty bi*ches	evil f*ckers
retarded fat	bum idiots	sh*ts annoying	damn sick
evil bastard	disgusting ni**as	bi*ches nasty	sick idiot
damn idiot	lil nasty	hospitals treating	brutal death
disgusting bi*ch	fools bi*ch	chaddi madness	timid piglets
bi*ch bad	feminist bi*ches	sick bastard	piglets cowering
nasty fat	evil bastards	damn idiots	fat ni**as
lil bi*ch	cruel bastard	fatuous idiot	hate hating
bad bi*ches	bi*ch annoying	bum bi*ch	damn mad
damn retarded	idiot retard	bird cheated	village idiots
damn annoying	ni**as hate	bomb suspect	redneck idiotic
sh*ts sick	ni**as sick	sick bi*ch	screaming bird
idiots imma	pig crap	damn sickening	disgusting terrorist
bi*ch hate	damn liar	bi*ch imma	delusional idiots
disgusting annoying	ni**as mad	bi*ches sick	shoots idiot
pathological liar	village idiot	threats fear	bi*ch shut
damn bad	damn evil	whine whine	insult idiots
worthless sh*ts	bi*ches bad	evil hell	fat shaming
hate bi*ches	nasty slut	evil evil	bi*ch mad
sh*ts mad	sh*ts nasty	evil damned	fat bi*ches
mad annoying	hurt idiots	damned evil	bi*ch steal
mad bi*ch	bad script	bi*ch retarded	
nasty bastard	hate bi*ch	presidential veto	
heartless bastard	bi*ches retarded	ni**as sickening	