Using a Three-step Social Media Similarity (TSMS) Mapping Method to Analyze Controversial Speech Relating to COVID-19 in Twitter Collections

1st Zhanyuan Yin*
Department of Mathematics and Department of Economics
University of California, Los Angeles
Los Angeles, CA, United States
yinzhanyuan1999@g.ucla.edu

2nd Lizhou Fan*
Program in Digital Humanities
University of California, Los Angeles
Los Angeles, CA, United States
lizhou@ucla.edu

3rd Huizi Yu
Department of Statistics and Department of Economics
University of California, Los Angeles
Los Angeles, CA, United States
huiziy@g.ucla.edu

4th Anne J. Gilliland
Department of Information Studies
University of California, Los Angeles
Los Angeles, CA, United States
gilliland@gseis.ucla.edu

***Abstract*—Addressing increasing calls to surface hidden and counter-narratives from within archival collections, this paper reports on a study that provides proof-of-concept of automatic methods that could be used on archived social media collections. Using a test collection of 3,457,434 unique tweets relating to COVID-19, China and Chinese people, it sought to identify instances of Hate Speech as well as hard-to-pinpoint trends in anti-Chinese racist sentiment. The study, part of a larger archival research effort investigating automatic methods for appraisal and description of very large digital archival collections, used a Three-step Social Media Similarity (TSMS) mapping method that aggregates hashtag mapping, TF-IDF Similarity Selection, and Emotion Similarity Calculation on the test collection. Compared to using a purely lexicon-based method to identify and analyze controversial speech, this method successfully expanded the amount of controversial contents detected from 21,050 tweets to 212,605, and the detection rate from 0.6% to 6.1%. We argue that the TSMS method could be similarly applied by archives in automatically identifying, analyzing, describing other controversial content on social media and in other rapidly evolving and complex contexts in order to increase public awareness and facilitate public policy responses.**

***Index Terms*—archival appraisal, archival description, COVID-19, social media, text mining**

I. Introduction

National libraries and academic special collections have been selectively collecting and preserving social media for almost two decades now. US President Donald Trump's notable use of Twitter to advance his presidential agenda over the past four years has drawn both public and government archives' attention to the ways in which social media can be used to disseminate materials that may have the status of official records and thus will need to be collected, managed and made available as such. Concerns about outside interference in US elections via social media has also increased scrutiny of the kinds of social media content, both overt and more subtle, that might be influencing public opinion. Furthermore, scholars in archival studies have been increasingly vocal about the need for archives to surface such latent narratives and hidden counter-narratives in support of calling out racism against and the marginalization of the experiences and voices of particular populations. However, even though controversial speech on social media has been recognized and for all these reasons has been a research focus since the beginning of the Web 2.0 era [2], identifying what to collect out of the vast volume created each day, and then rapidly describing its content and exposing racist and discriminatory trends, has proven to be difficult for the repositories that are engaged in these endeavors. Archives traditionally have been accustomed to manual processing of collections that often are acquired a significant amount of time after they were created. Two key issues that archives face in processing social media collections are how to generate nuanced description of extremely high volume, heterogenous and dynamic content that is archived in near real-time; and how to do so quickly enough that the resulting knowledge of the content can made available as quickly as possible to policy makers, educators and even social media platforms seeking to address problematic social concerns and trends such as growing racist sentiments.

The study reported in this paper is part of a larger research effort by a multidisciplinary team of academic that is addressing questions related to several aspects of archival practice:

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

The 1st and 2nd authors contributed equally to this work.
978-1-7281-6251-5/20/$31.00 ©2020 IEEE

archives to process the collections and make them available for research use. How can machine learning be used 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 [4]? As new content is added to the collection, can iterative application of those processes update, recontextualize and generate increasingly granular descriptions of the 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 making decisions about whether or not to ingest it into the archives? Could these processes be developed further as tools for reappraising archived digital collections?
4) *Archival transparency* – How can algorithmic reasoning be made transparent to archivists and users?
5) *Archival bias* – How can the reasoning of machine learning processes be assessed by humans, when should such assessment occur, and how can the reasoning be corrected if it is found to be developing 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?

During the COVID-19 pandemic, hate speech and language violence have been used on Twitter, particularly directed at China, Chinese people and people of Asian heritage [3]. With such a quickly unfolding situation and corresponding dynamics in public opinion, there is often limited time to develop complex language models that could be used to analyze and describe archived Twitter and other social media collections and use the results to combat problematic trends. Consequently, dictionary-based sentiment analysis and controversial content detection are among the most prevalent analytical methods being used [6]. We are proposing a more sensitive approach. Our study used the query "china+and+coronavirus" to scrape the Twitter API and obtain a collection of 3,457,434 unique tweets in English for analysis. Using purely dictionary-based methods, we identified 21,050 controversial tweets. However, the lexicon-based approach is inherently limited in identifying the overall structure and complex syntax of tweets. Furthermore, the massive volume of tweets and limited time that could be spent in annotating training data (controversial or not) led to many tweets being incorrectly marked as non-controversial. In order to respond more effectively to the rapidly evolving pandemic context, instead of having human annotators create a manually labelled dataset and then employing other supervised learning models to detect more controversial contents, we employed a Three-step Social Media Similarity (TSMS) mapping method for identifying and analyzing controversial speech.

II. Data Collection and Preparation

Since the discovery of pneumonia cases of unknown cause in Wuhan, China in December 2019, "China" and "Coronavirus" have been widely associated with each other in media reports and online discussion [12]. We queried the Twitter API from between January 31, 2020 and April 7, 2020, using the keywords "coronavirus+and+china" to build our archival text collection, the COVID-19 Hate Speech Twitter Archive (CHSTA).[1] To prepare the archived tweets for further analysis, we first removed stop words and punctuation while ignoring Twitter-specific non-textual strings such as the "#" in hashtags and URLs. We then applied lexicon-based methods for controversial tweet detection and aspect-based emotion scoring. We labeled tweets as "controversial" using a conservative definition, i.e., if they contain discriminatory words that are contained in the Hatebase dictionary [5]. For each tweets, we then generated emotion scores by tabulating the number of words related to each element (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust) in the emotion dictionary [8], [9]. We further normalized each emotion score by dividing it with by the tweet's word count.

III. Method: Three-step Social Media Similarity (TSMS) Mapping Algorithm

In fast-breaking social media, it is difficult to conduct sophisticated analyses of what is happening. If one instead were analyzing an historical dataset to understand the presence of controversial speech, one might approach the task differently. Therefore, to address the former circumstance, we propose an injective mapping

$$\mathbf{T}:\ \mathbf{M} \to \mathbf{N} \qquad n = T(m) \tag{1}$$

$$\text{subject to } m, n = \arg\min_{m,n} Dist_T(m, n)$$
$$m, n = \arg\min_{m,n} Dist_E(m, n)$$

in which we let $\mathbf{M} = \{\mathbf{0}, \mathbf{1}\}^{\mathbf{n_m} \times \mathbf{1}}$ stand for the space of non-controversial speech, and $\mathbf{N} = \{\mathbf{0}, \mathbf{1}\}^{\mathbf{n_n} \times \mathbf{1}}$ stand for the space of controversial speech. Here, n_m and n_n stand for the numbers of tweets in each non-controversial speech space and controversial speech space, respectively.

For $m \in \mathbf{M}$ and $n \in \mathbf{N}$, if the i^{th} tweet is chosen, then the i^{th} term of vector m, n is 1, and the rest are 0. Usually, m is a one-hot vector. $Dist_T(x, y)$ stands for topic similarity;

[1]The CHSTA Archive: https://github.com/lizhouf/CHSTA

$Dist_E(x, y)$ stands for emotion similarity. To derive the one-to-one mapping, we split the formation of the mapping into following three steps.

A. Hashtag Mapping

One major drawback of the sentiment analysis is topic irrelevance. For example, consider the following two statements: "I am happy that I passed the exam." and "I am happy that team A won the game." Although both of the statements express a positive (or more specifically, happy) emotion, because they are not talking about the same topic, we cannot regard them as "similar" statements.

Hashtags, the user-defined topic discourse of tweets, can be used to ensure that the range of tweet topics are relevant. Therefore, in our case, the first step is to find a mapping $\mathbf{T_1}$ such that

$$\mathbf{T_1} : \mathbf{M} \to \mathbf{N} \qquad [n]_i = H\left(\left[H_2 H_1^T m\right]_i\right) \tag{2}$$
$$H_1 = \{\mathbf{0}, \mathbf{1}\}^{\mathbf{n_m} \times \mathbf{n_k}} \qquad H_2 = \{\mathbf{0}, \mathbf{1}\}^{\mathbf{n_m} \times \mathbf{n_k}}$$

where $H(x)$ is the Heaviside step function. H_1 and H_2 stand for the hashtag matrices for controversial speech and non-controversial speech respectively, where n_k stands for the number of different hashtags in total, and the matrices are defined as:

$$[H]_{ij} = \begin{cases} 1 & j^{th} \text{ Hashtag exists in tweet } i \\ 0 & \text{else} \end{cases} \tag{3}$$

In other words, for each non-controversial tweet, the mapping $\mathbf{T_1}$ finds all candidate controversial tweets such that they have hashtags overlapping with controversial tweets.

B. TF-IDF Similarity Selection

For an instance of non-controversial speech m, after using mapping to find the index of all the instances of controversial speech n that have common hashtags, we can then find the indices of the common hashtags. These can be represented by

$$K(m, n) = (H_1 m) \cdot (H_2 n) \tag{4}$$

In the $n_k \times 1$ vector K, if hashtag k is a common hashtag then $[K]_k$ is equal to 1, otherwise it is 0. Next, we proceed to find out the second-step mapping for controversial speech instances that have the highest topic correlation with the non-controversial speech instances:

$$\mathbf{T_2} : \mathbf{N} \to \mathbf{N} \qquad n \to n_t \tag{5}$$

Since the topic similarity is determined by hashtags, the next step is to find out the levels of importance of different hashtags. For example, although "coronavirus" is a hashtag that is commonly matched and has high term frequency, it also has low specificity, which means that the hashtag may cover a wide range of related topics. Therefore, we may consider two tweets that have the specific hashtag "wuhan" in common to be more closely related than are two tweets that have the hashtags "coronavirus" or "china" in common. A traditional information retrieval method, Term Frequency–Inverse Document Frequency (TF-IDF), could balance between term frequency and specificity [13]. The formula to compute the TF-IDF score, defined as Score, for a hashtag k in non-controversial speech instance m is the following:

$$\text{Score}_{tf}(k, m) = \frac{f_{k,m}}{\sum_{k' \in m} f_{k'.m}} \tag{6}$$

$$\text{Score}_{idf}(k, m) = \log \frac{n_m}{n_k} \tag{7}$$

$$\text{Score}(k, m) = \text{Score}_{tf}(k, m)\text{Score}_{idf}(k, m) \tag{8}$$

where $f_{k,m}$ represents the frequency with which the hashtag appears among all hashtags in non-controversial speech instance m (i.e., we regard the list of hashtags of an instance of non-controversial speech as a document), usually equal to 1. $\sum_{k' \in m} f_{k',m}$, the sum of all hashtags in a hashtag list of non-controversial speech instance m, is usually the number of hashtags of tweet m. n_k stands for the number of non-controversial speech instances that contain hashtag k.

Inspired by the TF-IDF method, we can define the TF-IDF Similarity (Topic Similarity) of two tweets, m and the j^{th} controversial speech as [2]

$$Dist_T(m, n \cdot e_j) = \sum_{i=1}^{k} [K(m, n \cdot e_j)]_i \text{Score}(i, m) \tag{9}$$

It follows that the controversial speech instance that has the highest topic similarity can be expressed as (regarding m as a fixed non-controversial speech)

$$\mathbf{T_2} : [n_t]_j = \begin{cases} 1 & Dist_T(m, n \cdot e_j) = \\ & \min_i Dist_T(m, n \cdot e_i) \\ 0 & \text{else} \end{cases} \tag{10}$$

In other words, the topic similarity of two tweets (one non-controversial speech instance and one controversial speech instance) is the summation of the TF-IDF score with respect to all common hashtags in the non-controversial speech instance.

C. Emotion Similarity Calculation

Finally, we identify the controversial speech instance which has the highest emotional correlation by

$$\mathbf{T_3} : \mathbf{N} \to \mathbf{N} \qquad n_t \to n_s \tag{11}$$

The emotion of each tweet m is embedded into a length-eight vector representing eight emotions, E_m. The interpretation of the vector is the frequency of words representing different emotions appearing in the tweet, regularized by the length of tweet. Therefore, under the same topic, we could use a regular distance function, such as Euclidean distance, to measure the similarity of two vectors,

$$Dist_E(m, n \cdot e_j) = \sqrt{\sum_{i=1}^{8} (E_{m,i} - E_{n,i})^2} \tag{12}$$

[2]where $n \cdot e_j$ stands for picking the j^{th} controversial speech, e_j stands for the Standard Basis.

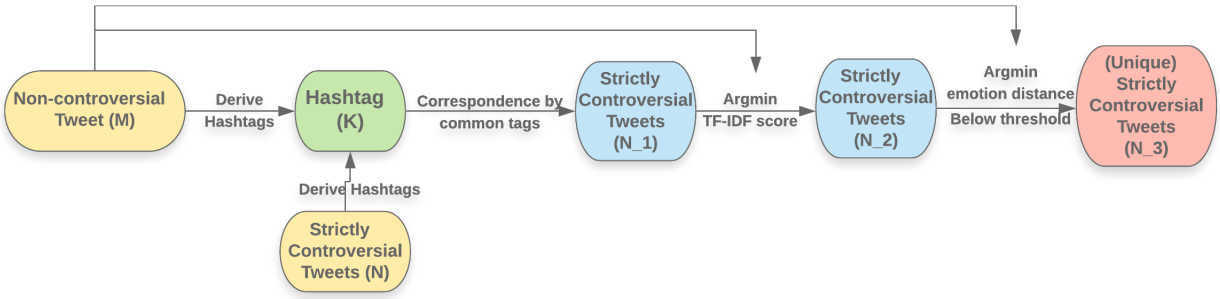

Fig. 1. Flow Chart of TSMS Mapping Method

As a result, the controversial speech that has the highest emotional similarity can be expressed as (regarding m as a fixed non-controversial speech)

$$\mathbf{T_3} : [n_e]_j = \begin{cases} 1 & Dist_E(m, n_t \cdot e_j) = \min_i \\ & (Dist_T(m, n_t \cdot e_i), \epsilon) \\ 0 & \text{else} \end{cases} \quad (13)$$

In equation 13 , ϵ is an activation threshold, i.e. only tweets that have emotional distance smaller than ϵ will be considered in the final mapping. Otherwise, the non-controversial tweet fails to be mapped with any controversial tweet.

The overall mapping $\mathbf{T}$, as represented in Figure 1, can be expressed as

$$\mathbf{T} : \mathbf{T_3} \circ \mathbf{T_2} \circ \mathbf{T_1} \quad (14)$$

D. Results

Our intention in building this TSMS mapping is to identify tweets related to COVID-19 that are implicitly controversial. When using lexicon-based methods to identify controversial speech, where there is a strong and obvious type of controversial content, researchers use a hard criterion, i.e., whether the tweet contains controversial words or not, to determine whether a tweet can be considered to be controversial speech. However, this approach excludes many tweets that are either politically or racially controversial but do not specifically contain controversial words. Using the TSMS mapping method, we successfully broaden how controversial speech can be understood and identified, as examples in Appendix A indicate.

Table I further summarizes our results: before the TSMS mapping, only 21,050 tweets are identified as controversial tweets, which is only 0.6% of tweets collected, while after the mapping, 191,555 additional non-controversial tweets can be mapped to one emotionally similar controversial tweet, and the fraction of controversial tweets reaches 6.14%.

Based on the successful mapping, we also discovered that for previously identified tweets (controversial contents defined by the lexicon-based method) and previously unidentified tweets, if they contain similar topics, they likely have emotional similarities. Therefore, it is feasible to detect whether two tweets contain similar context. As shown in Figure 2, intuitively, we can use the third quantile as the cut-off point

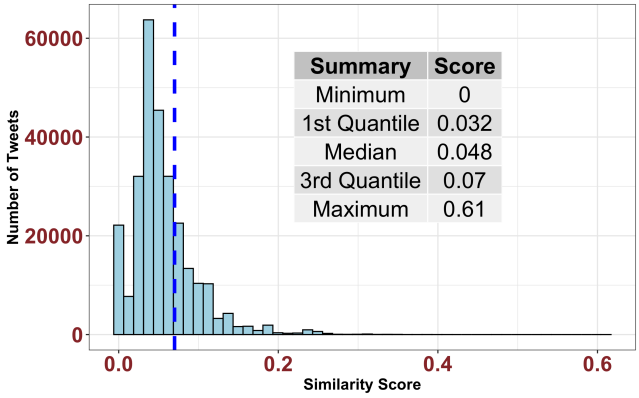

Fig. 2. Density Distribution of Similarity Score

Speech Type	Before	After
Controversial Tweet	21,050 (0.6%)	212,605 (6.1%)
Non-controversial Tweet	3,436,384 (99.3%)	3,244,829 (93.9%)

TABLE I
COUNT OF SPEECH TYPES BEFORE AND AFTER MAPPING

and thus 212,605 tweets remain identified as controversial tweets.

IV. Discussion

A. Related Work

Concurrent to our study, some other researchers have been trying to identify anti-Asian Hate Speech and Counter-hate Speech by formulating definitions and rules [14]. Pei and Mehta also use sentiment analysis to identify negative speech directed against Asians. They use hashtag information, but only racist hashtags, rather than all hashtags, are considered [11]. Topic modeling is another method that has been used for analyzing controversial terms related to COVID-19, but topic intractability would be less intuitive than using hashtags [1].

B. Advantages of the TSMS Algorithm

The three-step construction of the TSMS algorithm provides flexibility for model adjustment. In the second step, multiple

variations of TF-IDF [7] or similar term-importance measurements such as TF-RIDF or Mutual Information can be applied [10]. Similarly, in the third step, besides Euclidean distance, other distance measurements such as Hamming distance or cosine distance also apply. Thus, for this hybrid algorithm, we can choose alternative measurements in intermediate steps while keeping the body structure to obtain the best performance.

Another advantage of the algorithm is based on the assumption that since the distance function will return a continuous value, then for each instance of non-controversial speech, it can be mapped to a unique instance of controversial speech. Therefore, mapping $\mathbf{T}$ maps a one-hot vector to another and can be represented by a matrix M_T. Furthermore, the matrix M_T is sparse, which means that the complicated mapping could be easily stored and quickly manipulated.

C. Limitations and Future Work

To meet this study's goal of *rapidly* providing insight into how controversial speech is escalating on a social media platform, we are unable to implement too complicated a model or labor-intensive annotations. Therefore, complex learning models for sentiment analysis or controversial content detection are out of consideration, even though their precision could be higher. They might, however, be appropriate for analyzing collections of historical data and could be applied in a secondary archival description or analytical process at a later date.

For now, the TSMS model uses the third quantile cut-off point to detect and reduce the false positive controversial content. We plan to experiment on more hate-speech-related corpora and explore threshold selection criteria further.

Overall, however, we believe that the TSMS method of predicting, identifying and analyzing controversial content may present a generalizable approach that can be applied by archives and other repositories for rapid analysis and description of high volume social media content that has been archived in near real-time and where there are similar concerns about problematic trends, latent information, and submerged narratives, thus facilitating facilitating policymaking and educating the public regarding controversial content-related issues on social media. In further studies we will, therefore, be applying this method to archived collections on topical areas that share similarly rapidly evolving and complex dynamics.

REFERENCES

[1] L. Chen, T. Yang, J. Luo, H. Lyu, and Y. Wang. 2020. "In the Eyes of the Beholder: Sentiment and Topic Analyses on Social Media Use of Neutral and Controversial Terms for Covid-19." arXiv Preprint arXiv:2004.10225.

[2] D. DiNucci. 1999. "Fragmented Future." Print 53 (4). RC Publications Inc.: 32–33.

[3] L. Fan, H. Yu, Z. Yin. Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter. Proc Assoc Inf Sci Technol. 2020; 57:e313.

[4] W. Ernst and J. Parikka. 2013. Digital memory and the archive. University of Minnesota Press.

[5] Hatebase Inc. 2020. "Hatebase." https://hatebase. org/, Last accessed on 2020-04-26.

[6] H. Lyu, H. Lyu, L. Chen, and Y. Wang. 2020. "Sense and Sensibility: Characterizing Social Media Users Regarding the Use of Controversial Terms for Covid-19." arXiv Preprint arXiv:2004.06307.

[7] C. Manning, P. Raghavan, and H. Schütze. 2008. Introduction to Information Retrieval. Cambridge University Press.

[8] S. Mohammad and P. Turney. 2010. "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon." Association for Computational Linguistics, 26–34.

[9] S. Mohammad and P. Turney. 2013. "Crowdsourcing a Word–emotion Association Lexicon." Computational Intelligence 29 (3). Wiley Online Library: 436–65.

[10] C. Orasan, V. Pekar, and L. Hasler. 2004. "A Comparison of Summarisation Methods Based on Term Specificity Estimation." In LREC.

[11] X. Pei and D. Mehta. 2020. "#Coronavirus or #Chinesevirus?!: Understanding the Negative Sentiment Reflected in Tweets with Racist Hashtags Across the Development of Covid-19." arXiv:2005.08224v1.

[12] J. L. Ravelo and S. Jerving. 2020. "COVID-19 — a Timeline of the Coronavirus Outbreak." https://www.devex.com/news/ covid-19-a-timeline-of-the-coronavirus-outbreak-96396, Last accessed on 2020-04-26.

[13] R. Stephen. 2004. "Understanding Inverse Document Frequency: On Theoretical Arguments for Idf." Journal of Documentation 60 (5): 503–20.

[14] C. Ziems, B. He, S. Soni, and S. Kumar. 2020. "Racism Is a Virus: Anti-Asian Hate and Counterhate in Social Media During the Covid-19 Crisis." arXiv:2005.12423v1.

APPENDIX

Appendix A: Examples of mapping to conservative tweets

Original tweet: "WHY IS Cuomo COVID19 MAKING DEALS WITH China? ChinaLiedPeopleDied He is having Ventilators made and imported into NYC coronavirus ChinaLiedPeopleDied"

Mapped tweet: "The yellow race does not produce any art, science, or architecture. It wasn't until the yellow people came into contact with white culture and knowledge that they copied white science and architecture. covid19 coronavirus 2019nCoV china"

Similarity score: 0.054187527

Original tweet: "SARSCoV2 pandemic: Anger is growing at China over CoViD19 pandemic it caused and its apparent cover up attempt not to mention its ongoing data deceit and obfuscation"

Mapped tweet: "Fox fob off: seanhannity defending realDonaldTrump by blaming China and not the White House for America's lack of preparedness for COVID19: People all over the world suffering and dying as a result of China's lies."

Similarity score: 0.020579268

Original tweet: "Never discount just how stupid humans can be. To have a virus 100 due to China break out then have media freak out then stock market crash all to try to blame Trump is egregiously stupid. Twitter is another cog of the left and turno charges panic. Unreal coronavirus COVID2019"

Mapped tweet: "COVID2019 USA... Trump s overconfidence and confusion will make life miserable for his conservative GOP MAGA KAG Q QAnon tcot fans and may now get untold numbers killed off because he's an idiot. "

Similarity score: 0.049899553