Generating Connected Synthetic Electronic Health Records and Social Media Data for Modeling and Simulation

Anne M. Tall, Cliff C. Zou, and Jun Wang
University of Central Florida
Orlando, FL
anne.tall@knights.ucf.edu, czou@cs.ucf.edu, jun.wang@ucf.edu

ABSTRACT

Researchers are using big data sets, specifically large sets of electronic health records (EHR) and social media data, for experiments, such as investigating the potential to understand the spread of diseases and a variety of other issues. Applications of advanced algorithms, machine learning, and artificial intelligence indicate a potential for rapidly advancing improvements in public health. For example, several reports indicate that social media data can be used to predict disease outbreak and spread (Brown, 2015). Since real-world EHR data has complicated security and privacy issues preventing it from being widely used by researchers, there is a real need to synthetically generate EHR data that is realistic and representative. Current EHR generators, such as Synthea™ (Walonoski et al., 2017) only simulate and generate pure medical-related data. However, adding patients' social media data with their simulated EHR data would make combined data more comprehensive and realistic for healthcare research.

This paper presents a patients' social media data generator that complements and extends an EHR data generator. By adding coherent social media data to EHR data, a variety of issues can be examined for emerging interests, such as where a contagious patient may have been and others with whom they may have been in contact. Social media data, specifically Twitter data, is generated with phrases indicating the onset of symptoms corresponding to the synthetically generated EHR reports of simulated patients. This enables creation of an open data set that is scalable up to a big-data size, and is not subject to the security, privacy concerns, and restrictions of real healthcare data sets. This capability is important to the modeling and simulation community, such as scientists and epidemiologists who are developing algorithms to analyze the spread of diseases. It enables testing a variety of analytics without revealing real-world private patient information.

ABOUT THE AUTHORS

Anne M. Tall is a student in the Computer Engineering PhD program at University of Central Florida. She received a MSEE from Johns Hopkins University, Baltimore, MD and a BSEE from University of Maryland, College Park, MD. She is currently employed as a principal cybersecurity engineer at The MITRE Corporation. Her research focuses on computer and network security and systems engineering. The author's affiliation with The MITRE Corporation is provided for identification purposes only and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions or viewpoints expressed by the author.

Cliff C. Zou is an associate professor in the Department of Computer Science, University of Central Florida. He received a PhD from the Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA in 2005. His research interests include computer and network security, computer networking, and performance evaluation. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE).

Jun Wang is a full professor of computer science and engineering, and director of the Computer Architecture and Storage Systems (CASS) Laboratory at the University of Central Florida, Orlando, FL. He is recipient of the National Science Foundation Early Career Award 2009 and Department of Energy Early Career Principal Investigator Award 2005. He has authored over 120 publications in premier journals such as IEEE Transactions on Computers, IEEE Transactions on Parallel and Distributed Systems, and leading HPC and systems conferences such as HPDC, EuroSys, ICS, Middleware, FAST, IPDPS.

Generating Connected Synthetic Electronic Health Records and Social Media Data for Modeling and Simulation

Anne M. Tall, Cliff C. Zou, and Jun Wang
University of Central Florida
Orlando, FL
anne.tall@knights.ucf.edu, czou@cs.ucf.edu, jun.wang@ucf.edu

INTRODUCTION

This paper presents our strategies to generate synthetic social media data, that we have titled "SynSocial." It is based upon and linked to synthetic medical data generated by an Electronic Health Record (EHR) data generator, such as Synthea™(Walonoski et al., 2017). The approach to generating the data was designed as an initial open-source framework that could be expanded to generate social media data that is relevant and related to medical conditions and treatments over time. This type of generated social media data synthesized with medical healthcare data is needed for a variety of test and research applications, such as the detection and spread of diseases, early detection of illness or the effectiveness of behavior modification programs to improve health.

Like a recommender system in reverse, SynSocial generates social media data, (e.g., Tweets), based upon health care information. Previous research using recommender systems based upon real-world social media data has predicted a number of medical conditions, including flu outbreak trends (Brown, 2015), (Jordan et al., 2019), (Kahara, Haataja, & Toivanen, 2013), (Stromberg, 2013). Making connected synthetic social-media and healthcare data available to researchers enables the investigation of new algorithms and model development with open data that is free from security or proprietary controls. Current research is incorporating factors such as demographic and community information, such as age, sex, and relationships (friends, client/patient) in social media data as predictors of medical conditions (Geeta & Niyogi, 2016), (Singh, Dhawan & Pratibha, 2014), (Sadilek, 2012). The proposed synthetic social media data generator also incorporates these dimensions in the design. In this paper, we focus on Twitter data generation in this initial framework development.

Currently, several commercial entities are selling Twitter data analysis services. Twitter publishes some statistics about its usage. The motivation is to attract paid advertisers and not necessarily researchers, so the heuristic values are limited. However, Twitter provides an API and data can be scraped from this interface for a variety of research purposes. A challenge with using real-world social media data is the potential to disclose sensitive or personal information, especially when that data is connected with medical conditions.

Motivation

Public social media data sharing has been shown to be a new source of information to identify and analyze a variety of issues. Public health concerns and trends, in particular, have been identified by researchers as an area where new insights can be gained from social media data. Twitter is a leading source for this type of information, (Bucher, Christian & Meckel, 2013), (Schwartz, 2012), (Sneiderman, 2013), (Smith & Anderson, 2018).

However, to realize the potential of these insights, open, unsensitive information, free from Personally Identifiable Information (PII) is needed to develop algorithms and experiment with security features. For example, anonymization algorithms and residual risk of re-identification through inference could be tested using synthetic data. The challenge of developing synthetic social media data and anonymized data sets obtained through social media system APIs requires research, (Sagduyu, Grushin, & Shi, 2018). This project contributes to these efforts by proposing a method to generate synthetic social media data, specifically Twitter data, that is connected to synthetic medical data. These generated data sets can then be used to develop models and conduct analysis without concern about protecting PII and healthcare data as mandated by the Health Insurance Portability and Accountability Act (HIPAA), U.S. legislation that requires data privacy and security to safeguard certain medical information.

Novel Contributions

The unique and novel contribution for this data generator, SynSocial[1] is the connection of synthetic medical information to synthetically generated Twitter data. The medical information is connected to the social media data over a patient's lifetime. The generated messages are produced at a rate that is viewed as realistic based upon the age of the patient and their associated medical conditions. For example, the types of messages generated and the rate that they are generated vary over the lifetime of the patient. The data generator has been designed to enable adding higher levels of fidelity and realism as needed for various research objectives. The approach for validating the data produced by SynSocial was done by analyzing the frequency and quantity of messages when users are healthy (baseline) and when under medical conditions. This initial capability considers many factors influencing social media data generation and provides a framework that can be easily extended by others.

SYSTEM DESCRIPTION - DATA GENERATOR DESIGN

In this paper, we describe the design of SynSocial, our patient social media data generator, based on one particular EHR data generator. However, our design is generic and can be easily modified to be used with other EHR data generators.

The EHR data generator used, Synthea™, is an open-source synthetic EHR data generator that incorporates a wide variety of diseases (Walonoski et al., 2017)[2]. Initially, the top ten reasons to visit a Primary Care Physician (PCP) and the top ten diseases that cause loss of life, ("Two Top Tens"), were used to start the simulator project. The project has been expanded and currently has over 90 different modules that generate data on a wide variety of diseases.

The synthetic EHR include a number of states starting with "Condition Onset" and transitioning through various states associated with the disease progression, and then ending with "Terminal," states that result in generation of an Electronic Health Record (EHR). For SynSocial, the generation of medically related social media messages starts with condition-onset and ends at condition-abatement. If a medical condition end date is not included in the EHR data, an assumed date of one year after onset is used, based upon the idea that the number of social media messages written after a persistent long-term condition would drop-off after that period of time.

In SynSocial, birth and death dates are also used as input to the start and end of the social media message generation dates. For example, social media message generation starts at the age of 18. The number of messages generated decreases as the patients age and the contents of the messages corresponds to the patient's age group. Medical states such as prescribed medications and lab results included in the EHR are incorporated to expand the generation of medical-related social media messages.

Overall Design

The overall design of SynSocial is to enable the generation of a large volume of data in parallel. Each generated Tweet is appended as a JSON formatted message to the output file. Currently, modifications or deletions of previously generated data is not incorporated into the data generator design, but could be added in the future. Information such as marital status, education level, and other factors that influence a person's social media behavior could also be incorporated in the future. SynSocial uses a baseline Tweet generation rate based upon age and combines that with the generation of medical-related messages with the occurrence of a single or multiple conditions, as listed in the generated corresponding EHR data. The rate and contents of the generated Tweets is modified based upon combining and deconflicting the severity of the co-occurring conditions and baseline rate. Figure 1 highlights the program actions and interconnection. The logical flow of the SynSocial media data generator considers the message generation rate over the patient's lifetime and varies the rate and type of messages posted based upon the conditions and an age-associated baseline rate.

[1] Code is available at https://github.com/AnneMT/SynSocial
[2] Available at https://synthetichealth.github.io/synthea/

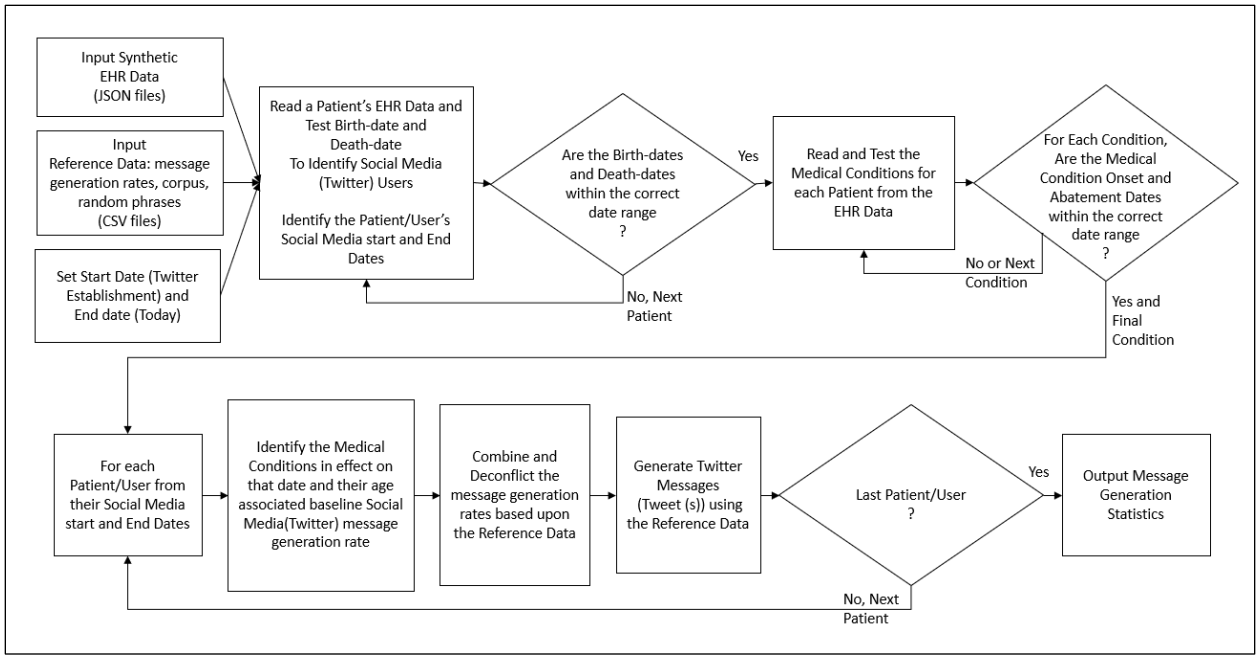

Figure 1. Overall SynSocial Social Media Data Generator Design

Data Input From Synthetic EHR Medical Data Generator

Synthetic medical data generators can be configured to produce a data set for a specified population size and support a variety of different configuration items, such as the locality of the population represented, and the format of the output data. SynSocial was designed to use as input JSON formatted files that contain each individual patient's medical condition over their lifetime.

An important field in the generated EHR data is the SNOMED-CT medical condition code. This is the standard language for encoding medical terms and conditions that has evolved over many years into international adoption. The translation of the codes into the medical condition is available online from SNOMED[3] and other sources such as the U.S. National Institute of Health (NIH)[4]. Short text messages based upon the SNOMED-CT codes are used to generate the full social media message (i.e., Tweet). For example, Table 1 lists example Tweets for the common disorder sinusitis. This corpus of message text is provided as a look-up file where the message is randomly selected from the phrases associated with the medical condition code. The original basis of the corpus is Twitter itself, however, none of the messages are an exact duplicate of real-world Twitter messages. The real-world user mentions and replies to real-world handles (i.e., use of "@Twitter username") has been removed and random phrases before and after message texts are added.

Table 1. Example Synthetic Social Media Contents

1	That was one of the worst cases of sinusitis I've had in a long time.
2	I have sinusitis. Any tips on getting rid of it? I would like to avoid antibiotics if possible.
3	What do you guys swear by for your allergies, esp. if you suffer from sinusitis?
4	Warm water Lemon, Ginger, Garlic, Turmeric and Cayenne pepper mix has helped me so much.
5	Sore eyes, tonsillitis and sinusitis, Wow
6	Currently suffering from sinusitis.
7	Can I get a new nose? This sinusitis got me good
8	Raging case of sinusitis.
9	Flu into a cold now acute sinusitis
10	Still feeling terrible and suffering from sinusitis

[3] https://www.snomed.org
[4] https://www.nlm.nih.gov/healthit/snomedct/

Data Output Format

The generated data is output in a JSON format based upon the messages that can be extracted from the Twitter developer interface API[5]. Not all fields are populated, however this can be expanded to incorporate additional complex dynamics in social media, (i.e., followers/following communities of interest). Tweets are the basic atomic building blocks that are posted, liked or reposted on Twitter. Tweets are also known as "status updates." The information contained in the tuple is based upon the information and format specified by the Twitter API. An example of the output social media message, in Twitter's Tweet JSON format is shown in Figure 2.


```
{"created_at": "Mon Jan 01 08:38:22 +0000 2007",
"id_str": "10101083822200710221562461",
"text": "absolutely, consulting a doctor regarding my chronic sinus condition, let's
chat later",
"user": {
        "id": 200511085042678197,
        "id_str": "200511085042678197",
        "name": "Tonja658",
        "screen_name": "@Tonja6fishnet",
        "location": null},
"place": {"country": "United States", "name": "Palmer Town, Massachusetts"},
"entities": {"hashtags": [], "urls": []},
"extended_entities": {"media": []}}
```


Figure 2. Example SynSocial Social Media Generated Message

Key fields generated by SynSocial are:

- Time stamp, (created_at), when the Tweet was published (created at date), with the time randomly generated
- Message content, (text), of the Tweet, randomly selected from the baseline age-correlated and medical condition corpus reference file
- Geographic location, (place), based upon the address in the EHR data file
- Author of the Tweet, (user screen_name), derived from the SynSocial generated name

This format includes the primary fields from a Twitter data object and represents data that is likely curated to analyze messages associated with medical conditions. A full Tweet contains additional fields that could be populated by SynSocial to meet a variety of research requirements. To match the emergence, use and enhancements of Twitter, the earliest date (time stamp) used is January 1, 2007 since Twitter didn't exist until 2006 and some features were created later, such as geo tagging.

Twitter Handle-Name Generation

SynSocial uses the first name generated by the EHR data generator combined with a random word (noun) appended to create a Twitter nickname or handle. The random word is currently chosen from a reference file, however, to increase the diversity, the words could also be created using a random word or name generation tool[6]. The desired level of realism would influence the source (dictionary or random name generation tool).

Twitter Message Generation Model

Our SynSocial generator creates two types of Twitter messages (Tweets): normal messages, and medical-condition messages. Normal messages are the general tweet messages generated by users that are not related to their medical conditions; medical-condition messages are tweet messages where the users talk about or discuss their current medical conditions and health concerns.

For each type of Twitter messages, the number of messages generated by a user per day will be modeled to follow Poisson Distribution[7] $X \sim Pois(\lambda)$ where the rate λ is the mean value of the number of messages generated per day. At any given time, a user may be healthy, or may have one or more medical conditions. Let us denote the number of Twitter messages generated by a particular user in a day is N, then:

[5] https://developer.twitter.com/en/docs/tweets/data-dictionary/guides/tweet-timeline
[6] https://randomwordgenerator.com/name.php
[7] https://en.wikipedia.org/wiki/Poisson_distribution

$$N = \alpha \cdot N_h + N_m \quad (1)$$

Where N_h is the number of normal Twitter messages when the user is healthy (called 'baseline' messages) and N_m is the number of medical-condition Twitter messages. $N_h \sim Pois(\gamma)$ where γ is the rate of baseline messages. $N_m \sim Pois(\lambda)$ where λ is the rate medical-condition Twitter messages are generated. When the user is healthy without any medical conditions, λ would be 0.

In Equation (1), the important parameter, α ($\alpha \in [0,1]$), represents the *illness impact* to a user's daily normal Twitter message generation: when a user is sick with one or multiple medical conditions, the user would reduce their normal Tweets, but will generate some messages related to the illness, expressing their feelings, comments, or concerns towards their current medical conditions. If a user is seriously sick, such as staying in hospital, α could be as small as 0 meaning that the user has no ability to generate normal Twitter messages due to this medical condition.

Suppose there are *n* medical conditions in the generated EHR data. For each medical condition SNOMED-CT code $i, (i = 1, 2, \cdots n)$, we define in SynSocial the corresponding illness impact factor α_i, and medical-condition Tweet Poisson distribution rate λ_i. If the user has one and only one medical condition i , the user's generated medical-condition tweet rate is simply $\lambda = \lambda_i$, and α in Equation (1) is simply equal to α_i.

If the user has multiple illnesses, (i.e., the user has a set of medical conditions $\mathcal{S}$), the user's daily generated Twitter messages would be modeled by Equation (1) with the following parameters:

$$\alpha = \min_{i \in \mathcal{S}} \alpha_i \quad \text{and } \lambda = \sum_{i \in \mathcal{S}} \lambda_i.$$

Table 2 shows the parameters used for the normal social Tweet messages generated by our SynSocial. For normal Tweet message generation, we classify users based on their age group. γ is the Poisson distribution rate, i.e., the average number of normal Tweet messages generated by a user per day. The daily tweet rate shown in Table 2 are set in an Excel configuration file in SynSocial; so they can easily be changed to support different research objectives. Example topics contained in the corpus of message texts are also listed. The actual message is randomly selected from an Excel file and then a random phrase is appended to the beginning and end of the message.

Table 2. Baseline Message Generation Rate

Age (years)	Daily Normal Tweet Rate (γ)	Example Tweet Text Topics
Birth to 18	0	nil
18 to 20	4	college, dating, job search, first job, working, wedding, first home
20 to 25	4	college, dating, job search, first job, working, wedding, first home
25 to 30	4	parties, children child care, food, vacation, sports, job change
30 to 40	3	children, moving, job, promotion, social activities
40 to 50	3	food, vacation, sports, hobbies
50 to 60	2	grown children, grandchildren, moving, job, social activities
60 to 70	2	retirement, vacations, home repair, hobbies
70 to 85	1	retirement, travel, home repair, hobbies
85 to 99	1	travel, home repair, hobbies
99 or older	0	nil

SynSocial is flexible and open so that alternative approaches to creating the corpus of message text can be used. For example, a random text or phrase generator[8] could be used to create the message corpus.

Example values of the medical-condition Tweet generation parameters, λ and α, for several SNOMED-CT codes is listed in Table 3.

[8] such as http://theidiomatic.com/

Table 3. Example Medical Condition Message Generation Rates and Severity

SNOMED-CT Code	Condition	Daily Medical Condition Message Generation Rate (λ)	Impact on the Baseline Rate (α)
4448144009	Viral Sinusitis	1	0.25
162864005	Obesity	2	1
40275004	Contact Dermatitis	1	0.5
72892002	Normal Pregnancy	2	0.5
198992004	Antepartum Eclampsia	1	0

For this example, the medical condition code 198992004, Antepartum Eclampsia, is considered a severe medical condition, which would dramatically impact the patient's normal social media post activity, making the normal post to be zero. In such situation, the patient would be posting a small number of messages only about their severe conditions and not about other topics.

DATA GENERATOR EVALUATION

Evaluation of the data produced by SynSocial was considered based upon the quantity and quality of the data generated. The publicly available statistics on the number of real-world Twitter messages generated by age and medical conditions were identified. For example, the prevalence of illness related Twitter messages has been identified by researchers (Jordan et. al, 2019). In Figure 3, an example of the number of generated baseline and medical condition social media messages is shown over multiple years. The occurrence of medical conditions is indicated, with the impact of a severe medical condition, Antepartum Enclampsia, on the suppression of baseline messages highlighted. Between 2016 and 2017 this synthetic patient did not have any medical conditions, and the messages are generated at a rate that matches a Poisson Distribution for a mean of 3 which corresponds to the age bracket (30 to 40) of this patient during this time, as listed in Table 2. The 2011 to 2016 timeframe was omitted (as noted by the jagged line) to show more of the medical timeline for this example social media user/patient.

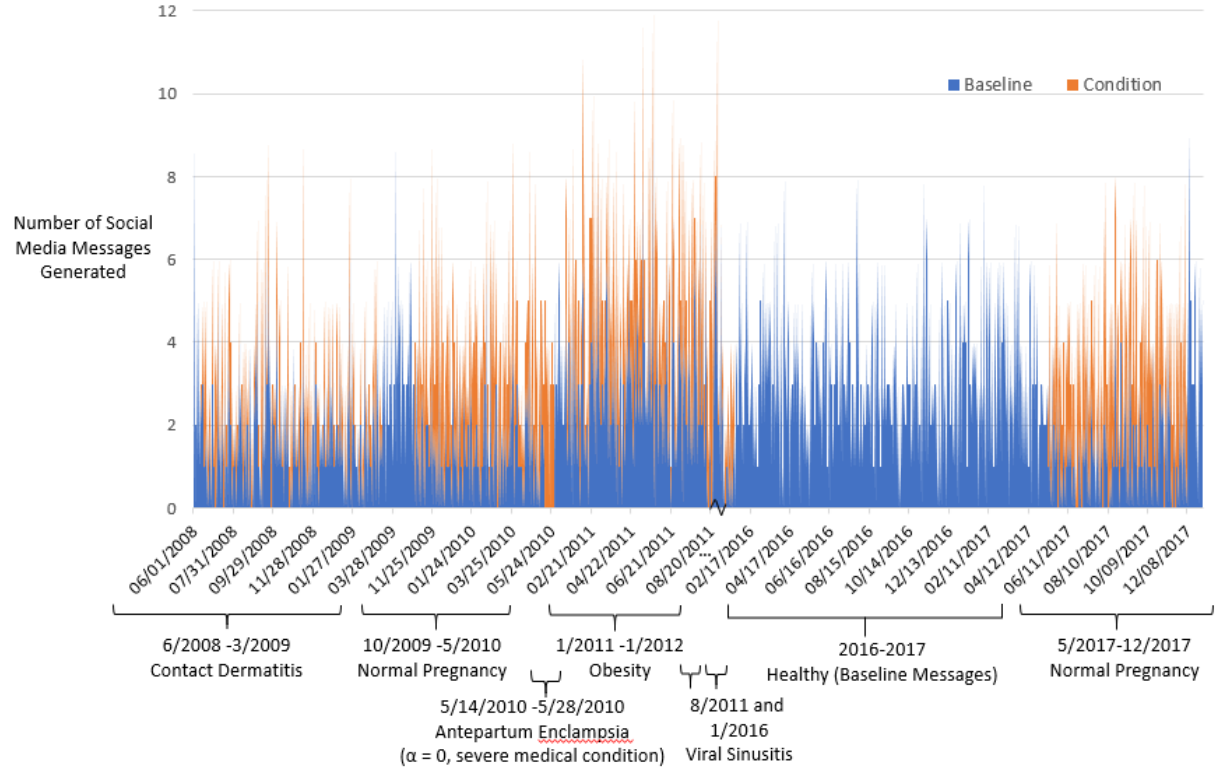

Figure 3. Example Social Media Data Generation for an Individual User/Patient

To examine the qualitative aspects of the generated messages, whether the contents of the messages correspond to what people Tweet about was considered. Less is published about trends in message contents as it relates to medical

conditions. The overall goal was to ensure that the generated data contained no confidential or sensitive data. The message body corpus was based upon real-world data, however Twitter "mentions" which reference real user names were removed and additional random data was added to generate synthetic data. The process of extracting a baseline and medical condition corpus of messages could be further automated to increase the anonymization, scrambling of a larger set of real-world messages used as the base for the message generation. The size of the generated data is summarized in Table 4 below. This makes this synthetic data well suited for big data experiments that require representative data that has both a high degree of sensitivity, (such as EHR data), and open unclassified information, (such as social media data).

Table 4. Example Social Media Data Generation Size

Input Patient's EHR Data File Size	Number of Medical Conditions in Input Data	Output User's Social Media Message File Size	Number of Medical Condition Messages (Tweets)	Number of Baseline Messages (Tweets)
718 KB	10	7,413 KB	1,791	12,818

There are several research projects (Xia et al., 2014) (Yu et al., 2014) that focus on understanding the relationships between social media users, e.g, who is following who, numbers of likes. This type of research could be incorporated into SynSocial as mentions (e.g., using an "@" tag). Also, there is research (Rosenthal et al., 2019) associated with understanding the sentiment of messages which could be used in building the corpus of the referenced message sets. Artificial generation of message text to portray conversations is also an area of research that could influence the design, however the nature of Twitter, is more akin to a micro-blog post rather than a conversation, so a Twitter chat-bot[9] might be applied to represent comments on messages. These are areas for potential further research, development and expansion of SynSocial.

RELATED WORK

Researchers have proposed several alternatives for social media data generation. Specifically, Yu et al. proposed the BSMA-GEN simulation, (Yu et al. 2014). This effort addressed the need for parallel execution and scaling to produce a large data set. The contributions also included ensuring the produced data is in a realistic format and addresses behaviors such as re-Tweeting.

Sagduyu, Grushin and Shi proposed a synthetic social media data generator that uses a novel concept in generating synthetic graphs to realistically address who is talking to whom (Sagduyu, Grushin, & Shi, 2018). However, a challenge in applying this concept to Twitter, is that the media operates as a broadcast to many followers rather than a direct person to person exchange. This research effort also addressed synthetic text generation using innovative approaches such as chat-bots or social media bots. They qualified the utility of this strategy through human experiments to measure the realism of the synthetically generated messages. Overall, they were able to produce texts that are grammatically correct and coherent.

Another important area that has been researched is geographic tag (geo-tag) references in messages and their relevance in studying various issues. For example, Sadilek, Kautz, and Silenzio examined disease transmission using a combination of social media posts and associated geo-tag information, (Sadikek, Kautz, & Silenzio, 2012). Moreira, Tiago, and Pianho used geo-tags in combination with social media data to examine emotions and stress in smart cities, (Moreira De Oliveria & Pianho, 2015). Indicators of mental health issues in social media data, as proposed by Yazdavar et al., is an interesting emerging area that could provide great insights, but also contain many PII and HIPAA sensitivities, (Yazdavar et al., 2018). Nugyen et al. all examined food-related illnesses using social media posts and associated geo-tag data, (Nguyen et al., 2017). These innovative research efforts are providing opportunities to gain greater insights into a number of health issues based upon real-world data sets. Synthea has been recently updated to include COVID-19 medical conditions, so testing algorithms that track their geographic area using geo-tagged Tweets might be an interesting area for investigation.

To further these types of research efforts, models and simulations can be developed, tested and enhanced using synthetically generated social media data connected to healthcare information as proposed in this paper. Future efforts to advance the initial concepts proposed would be to conduct validation and verification of the generated messages

[9] https://marketing.twitter.com/emea/en_gb/insights/how-to-plan-and-analyse-a-twitter-chatbot

for frequency and relevance against actual real-world data, (i.e., compare the generated Tweets to real-world Tweets). The realism could be enhanced using additional rich formats and data types, (e.g., pictures, web-site links, hashtags, mentions, images, videos, clips, and sound). Another area for future investigation could be modeling the behavior and impact of social events and influencers as input to text generation. This may enable the analysis of using social media to encourage more healthy behaviors, such as exercising and eating healthy foods.

CONCLUSION AND FUTURE WORK

In this paper we proposed SynSocial social message data generator that is connected to synthetically generated medical data from an open source medical data generator. SynSocial provides a useful resource for developers experimenting with new insights that can be gained from social media data without concerns for PII and HIPAA requirements. The initial framework of the design allows for it to be extended for higher fidelity realism as needed for a larger number of complex medical conditions. This is a first, unique effort to provide the tools necessary to advance large scale data analysis from two previously unconnected sources, one very sensitive (healthcare data) and the other open public (social media data), thus enabling the development of new data analysis capabilities.

ACKNOWLEGEMENTS

This work was supported in part by NSF grants DGE-1915780. Any opinions, findings, conclusions, and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agency.

REFERENCES

Brown, J. (2015, January 8). Using social media data to identify outbreaks and control disease, *Government Technology-Emergency Management-Public Health*. Retrieved from: http://www.govtech.com/em/health/Social-Media-Data-Identify-Outbreaks.html

Bucher, E., Fieseler, C., & Meckel, M. (2013, January 7). *Beyond demographics - explaining diversity in organizational Social Media Usage,* 2013 46th Hawaii International Conference on System Sciences, Maui, Hawaii. Retrieved from: https://ieeexplore.ieee.org/document/6480388

Geeta & Niyogi, R. (2016, September 21). *Demographic analysis of twitter users,* 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India. Retrieved from: https://ieeexplore.ieee.org/document/7732461

Jordan, S. E., Hovet, S. E., Chun-Hai Fung, I., Liang, H., Fu, K., & Tsz Ho Tse, Z. (2019). Using twitter for public health surveillance from monitoring and predictions to public response, *Data*, 4(6), doi: 10.3390/data4010006. Retrieved from: https://www.mdpi.com/2306-5729/4/1/6

Kahara, T., Haataja, K., & Toivanen, P. (2013, December 4). *A novel recommendation system approach utilizing social network profiles,* IEEE 13th International Conference on Hybrid Intelligent Systems (HIS 2013), Tunis, Tunisia. Retrieved from: https://ieeexplore.ieee.org/document/6920474

Moreira De Oliveria, T., & Painho, M. (2015, June 17). *Emotion & stress mapping: Assembling an ambient geographic information-based methodology in order to understand smart cities*, 2015 10th Iberian Conference on Information Systems and Technologies (CISTI), Aveiro, Portugal. Retrieved from: https://ieeexplore.ieee.org/abstract/document/7170469

Nguyen, Q. C., Meng, H., Li, D., Kath, S., McCullough, M., Paul, D., Kanokvimankul, P., Nguyen, T.X., & Li, F. (2017, May 4). Social media indicators of the food environment and state health outcomes, *Public Health*, Elsevier. Retrieved from: https://www.ncbi.nlm.nih.gov/pubmed/28478354

Rosenthal, S., Mohammad, S., Nakov, P., Ritter, A., Kiritchenko, S., & Stoyanov, V. (2019, December 5*). SemEval-2015 task 10: Sentiment analysis in twitter,* SemEval-2015. Retrieved from: https://arxiv.org/abs/1912.02387

Sadilek, A. (2012). *Modeling human behavior at a large scale*, (Doctoral dissertation), University of Rochester. Retrieved from: https://dl.acm.org/citation.cfm?id=2519086

Sadilek, A., Kautz, H., & Silenzio, V. (2012, July 22). *Predicting disease transmission from geo-tagged micro-blog data,* Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, pages 136-142. Retrieved from: https://dl.acm.org/citation.cfm?id=2900728.2900748

Sagduyu, Y., Grushin, A., & Shi, Y. (2018, August 16). Synthetic social media data generation, *IEEE Transactions on Computational Social Systems*, 5(3). Retrieved from: https://ieeexplore.ieee.org/document/8438309

Schwartz, A. (2012, August 2). Twitter knows when you'll get sick before you do, *Fast Company*. Retrieved from: https://www.fastcompany.com/1680262/twitter-knows-when-youll-get-sick-before-you-do

Singh, K., Sanjeev D., & Pratibha. (2014, September 25). *Real-time data elicitation from twitter: Evaluation and depiction strategies of tweets concerned to the blazing issues through twitter application,* 2014 5th International Conference - Confluence The Next Generation Information Technology Summit, Uttar Pradesh, Noida, India. Retrieved from: https://ieeexplore.ieee.org/document/6949269

Smith, A. & Anderson, M. (2018, March 1). Social media use in 2018, *Pew Research Center.* Retrieved from: https://www.pewinternet.org/2018/03/01/social-media-use-in-2018/

Sneiderman, P. (2013, January 24). Using twitter to track the flu: Researchers find a better way to screen the tweets, *Johns Hopkins University*. Retrieved from: http://releases.jhu.edu/2013/01/24/using-twitter-to-track-the-flu/

Stromberg, Joseph. (2013, November 8). Your tweets can predict when you'll get the flu, *Smithsonian.com*. Retrieved from: https://www.smithsonianmag.com/science-nature/your-tweets-can-predict-when-youll-get-the-flu-180947646/

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2017, August 30). Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, *Journal of the American Medical Informatics Association*, 25(3), 230–238. Retrieved from: https://doi.org/10.1093/jamia/ocx079

Xia, F., Li, Y., Yu, C., Ma, H., & Qian, W. (2014, June 4*). BSMA: a benchmark for analytical queries over social media data*. Proceedings of the VLDB Endowment. Retrieved from: https://doi.org/10.14778/2733004.2733033

Yazdavar, A. H., Mahdavinejad, M. S., Bajaj, G., Thirunarayan, K., Pathak, J. & Sheth, A. (2018, June 4*). Mental health analysis via social media data*, 2018 IEEE International Conference on Healthcare Informatics (ICHI), New York, NY. Retrieved from: https://ieeexplore.ieee.org/document/8419435

Yu, C., Xia, F., Zhang, Q., Ma, H., Qian, W., Zhou, M., Jin, C., & Zhou, A. (2014). *BSMA-GEN: A parallel synthetic data generator for social media timeline structures,* DASFAA 2014, Part II, LNCS 8422, Springer International. Retrieved from: https://link.springer.com/chapter/10.1007%2F978-3-319-05813-9_40