Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Signature:

Yunjie Wu April 1, 2023

ClusT: Interactive Visualization Tool for Deep Constrained Clustering on Tweets

By

Yunjie Wu

Joyce C. Ho, Ph.D.
Advisor

Emily Wall, Ph.D.
Co-Advisor

Computer Science

Joyce C. Ho, Ph.D.
Advisor

Emily Wall, Ph.D.
Co-Advisor

Zachary Binney, Ph.D.
Committee Member

Shivani A. Patel, Ph.D.
Committee Member

2023

ClusT: Interactive Visualization Tool for Deep Constrained Clustering on Tweets

By

Yunjie Wu

Joyce C. Ho, Ph.D.
Advisor

Emily Wall, Ph.D.
Co-Advisor

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences of
Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Computer Science

2023

Abstract

ClusT: Interactive Visualization Tool for Deep Constrained Clustering on Tweets
By Yunjie Wu

Healthcare indices are used to evaluate the overall accessibility of public health resources for a given region and have been found to be especially useful for predicting post-treatment outcomes for a variety of diseases. Heart failure (HF) patients' outcome after receiving treatment, for instance, has been discovered to be strongly correlated with the patient's residing neighborhood, a variable yet to be incorporated into the process of health indices' generation. Based on preliminary research that shows census-level Twitter data can be utilized to capture neighborhood impact, we introduce a visual analytic system that enables the iterative refinement of this new data. Specifically, the system enables machine learning experts, healthcare professionals, and policymakers to (1) preprocess tweets retrieved by specified keywords, e.g., eliminating URLs, stemming words, etc., (2) extract keywords and their corresponding embeddings, and (3) customize, inspect, and refine a topic model through interactive clustering. Ultimately, data produced by this system, along with other data sources, allows for the refinement of a more environment-reflective healthcare index.

ClusT: Interactive Visualization Tool for Deep Constrained Clustering on Tweets

By

Yunjie Wu

Joyce C. Ho, Ph.D.
Advisor

Emily Wall, Ph.D.
Co-Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Computer Science

2023

Acknowledgments

I would like to express my greatest gratitude to my advisors, Dr. Joyce C. Ho and Dr. Emily Wall, for providing me with this inspirational research idea. I sincerely appreciate their guidance and encouragement throughout the course of my thesis work, as well as their patience and all hands-on advice. Without their mentorship, this thesis would not have been possible.

I am grateful to Dr. Jing Zhang for his critical contributions to my research. His insights and suggestions played a vital role in shaping my project. I also want to thank Dr. Jinho Choi and my committee members, Dr. Zachary Binney and Dr. Shivani A. Patel, for their valuable feedback, which helped to improve my thesis.

I am fortunate to have had the unwavering support of my family during my academic pursuits. Their love and encouragement carried me through the ups and downs of my journey. I also wish to express my appreciation to my friend Alexandra (Xinran) Li for her help during my studies and for being my companion throughout the college years.

Finally, I would like to thank all the professors I have encountered during my time at Emory University. Their kindness and support have made a lasting impact on me. I am grateful for the opportunity to undertake my undergraduate studies at Emory University.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

It has been shown that the neighborhood in which individuals have resided for a certain amount of time has a substantial impact on their overall health [11, 37]. The availability of social and economic resources in these communities greatly influences the health outcomes of patients with various diseases. As motivating examples, availability and access to healthy foods, healthcare services, and places to exercise have shown to be associated with positive health outcomes [17, 22]. Neighborhood aesthetics were shown to have a negative effect on glycemic control [34, 35] while social cohesion can have a positive association [4]. Predictions such as the rate of relapse, readmission, and death are valuable information for health professionals to devise more effective and customized treatment strategies for patients according to the neighborhood they are from. Therefore, previous research has attempted to establish a connection between patients' neighborhood environment and their post-treatment outcomes. To accomplish this goal, researchers found it crucial to devise a method to characterize the social and economic status of the neighborhood.

Existing measures such as Social Deprivation Index (SDI) [8] and Area Deprivation Index (ADI) [32] capture demographic features, such as annual income, education level, and non-employment rate of a neighborhood environment at several levels of

geological units collected from the US Census American Community Survey (ACS) [7]. Using SDI at the census level to examine its relationship with the post-treatment outcome of heart failure (HF) patients specifically revealed a negative association [26]. That is, the higher the SDI, the worse the post-treatment outcome for HF patients.

Given the criticality of healthcare measures such as SDI for characterizing a neighborhood's healthcare resources and health outcomes, distributing healthcare resources, and informing decision-making for families on where to live, it warrants further investigation to improve upon such metrics. For instance, given that SDI is derived from census data, it has limited localization to pre-defined census tracts, which have numerous limitations, among which is that it is updated only once per decade and may not be fine-grained enough to reflect neighborhood-level differences.

Hence, there is a need for an accurate depiction of the neighborhood environment, as existing measures fail to capture various important characteristics such as health-related resources and food deserts. There is also a need for a data source that can provide near real-time assessments of the environment and does not require additional collection. One particularly promising potential source of data is Twitter, which is a rich untapped source of data that can provide finer-grained information, with tweets sent from a specific location indicative of the availability of local health resources.

Preliminary Work Our preliminary research using 17 census tracts in the Atlanta metropolitan area indicated that census-level Twitter data could be utilized to reveal neighborhood influences [42]. We demonstrate Twitter, Google Places, and Foursquare can be used to construct a neighborhood environment index across 6 different categories: Park, Pharmacy, Grocery, Restaurants, Sports/Recreation, and Health. Table 1.1 summarizes the prediction performance of our health index compared to SDI on predicting 30-day readmission for 804 patients with HF. As can be seen, the results suggest the efficacy of leveraging information from these novel data

streams to provide a neighborhood assessment.

Table 1.1: Model performance of novel environmental measures and the traditional social deprivation index for predicting 30-day heart failure readmission in 804 patients at Emory Healthcare. A higher area under the receiver operating characteristic curve (AUC) denotes better prediction performance.[42]

Neighborhood measure	**AUC**
Park	0.519
Pharmacy	0.505
Grocery	0.517
Restaurants	0.532
Sports/Recreation	0.502
Health	0.516
SDI	0.514

Analysis of the Twitter data uncovered a potential limitation of our approach for approximating the neighborhood environment. Our initial analysis simply counted the number of geo-tagged Tweets with specific keywords. For example, to assess health resources, we used the keywords hospital, doctor, and dentist. However, as shown in Figure 1.1, this meant that two tweets with different topics are aggregated together and considered equal. Thus, there is a need to capture the topic heterogeneity within the collected tweets. Yet extracting information from the original tweets and identifying import topics requires significant data processing and domain expertise in natural language processing (NLP) and machine learning (ML).

Figure 1.1: Example tweets from our dataset, wherein the upper tweet pertains to the subject of cancer while the lower tweet pertains to the topic of health insurance.

In this thesis, we propose ClusT, an innovative, visual analytic system that enables non-experts in NLP and ML such as public health professionals, and policymakers to participate in the refinement and topic-building process, ultimately leading to a more neighborhood-reflective index.

Our system consists of three distinct phases: unstructured text preprocessing, keyword representation extraction, and interactive clustering. The preprocessing phase allows users to leverage their own judgment and preferences to remove redundant and non-informative information. Furthermore, we design the visualization to assist users in understanding how data is processed prior to the keyword phase. Our system's keyword phase primarily takes place in the back end, where we extract keywords from preprocessed tweets and generate their corresponding word embeddings for clustering. In the interactive clustering stage, our system asks users to identify keywords that they feel should belong to the same or different cluster. Through iterative refinement based on user input, the system creates a better topic model, which is then used to enhance Twitter data. We demonstrate the use of ClusT to develop a better measure of local health resources by refining Twitter data associated with the health category.

With a more localized healthcare index, we hope to establish a more direct relationship between the neighborhood deprivation of HF patients and their treatment outcome, allowing for the development of more effective treatment for HF patients.

Chapter 2

Background and Related Work

2.1 Neighborhood Environment & Patient health outcome

The term neighborhood environment refers to the socioeconomic status of a geological region. Census data has frequently been used to assess the socioeconomic status of a community based on factors such as the unemployment rate, income, occupation, housing, education levels, etc. [24]. These factors are associated with whether people have access to the resources necessary to live a healthy lifestyle, including public facilities that promote a healthier lifestyle, economic conditions that allow for health-related spending and healthcare, and hospitals, clinics, and health services that provide direct access to health checks and treatment. SDI and ADI are well-established indicators of the socioeconomic condition of families in a specific neighborhood, composed of several demographic characteristics [8, 32]. Its data are divided by geological units, including the census tract proxy. Census tracts are subdivisions of counties and have, in general, homogeneous populations with similar socioeconomic levels and living criteria [29]. Therefore, SDI and ADI have been used by previous research as a healthcare index to predict patient outcomes [26, 28].

An alternative index, "neighborhood deprivation," which measures residents' accessibility to public health resources, has been proposed in an earlier study to reflect the health-related access and outcome of neighborhoods [24]. This new neighborhood deprivation index (NDI) is constructed through a principal analysis of census data focusing on public-health-related indices and is associated with adverse health outcomes [24, 31]. For HF specifically, there has also been previous research that has found low-income populations in the southeastern states of the United States to have a higher chance of developing heart failure [1].

Information contained in the census data from which SDI, ADI, and NDI are derived, however, is not always up-to-date due to the relatively low frequency of its data collection and its limited refinement capacity. To tackle this problem, we retrieve Twitter data based on the census tract the tweets were posted from.

2.2 Preprocessing

Text preprocessing is essential for any practical use of unstructured data and any application of text mining techniques, including classification, clustering, and visualization [38]. Numerous works have studied the effectiveness of text preprocessing techniques. According to one study, the preprocessing phase consists primarily of three steps: stop-word elimination, stemming, and term frequency-inverse document frequency (TF-IDF) [40]. This paper examines the importance of stop-word removal, which entails removing articles, prepositions, and pronouns, as well as the differences between various types of stemming models and a brief introduction to TF-IDF [40]. The majority of text-preprocessing-related studies, however, omit components that could aid users' comprehension of the data preprocessing procedure and interactivity that would allow users to customize their own NLP algorithms and functions. Our system implements independent tabs for preprocessing usage in which customization

of NLP methods for users without NLP expertise is supported.

Specifically, our preprocessing tab has a framework that adheres to the general guideline presented by Vijayarani *et al.* but also supports user customization facilitated by visualizations [40]. Due to the fact that the most appropriate preprocessing techniques are highly dependent on the text data on which they are performed, there is also a need to make modifications to account for the Twitter data's unique characteristics [15]. Earlier studies have also specified structural elements that require elimination, like URLs and hashtags [2]. By employing both conventional preprocessing techniques and user interactivity, our objective is to prepare the data for subsequent word embedding extraction and clustering.

2.3 Interactive Clustering

Interactive clustering is an approach that integrates human expertise with machine learning algorithms to improve the clustering process. This method is particularly useful when dealing with complex or noisy data, where traditional unsupervised clustering techniques may fail to yield satisfactory outcomes. By involving human input, interactive clustering can exploit domain knowledge and facilitate a better understanding of the underlying data structure, leading to more accurate and interpretable cluster assignments [16].

User interactions within topic modeling systems can be classified into three categories based on the unit of interaction: word-level, document-level, and topic-level. Focusing specifically on word-level interactions, users can manipulate topics by dragging words between them to split, combine, or add topics [36], defining pairwise constraints [5], and employing other techniques.

Our system employs a 2D main view, as used by Topiclens and various other systems, which has shown to be more accessible and easier for users to comprehend and

engage with compared to 3D plots [20]. To better support domain experts in comprehending the data and providing constraints, it is advantageous to supply contextual information within the system [19]. In light of these considerations, our approach to interactive clustering aims to facilitate user engagement by focusing on word-level interactions and presenting the data in an easily understandable 2D view. Furthermore, we strive to enhance the user experience by offering contextual information that can aid domain experts in grasping the data's intricacies and providing relevant constraints. By addressing these aspects, our system seeks to promote more accurate and interpretable cluster assignments, thus improving the overall clustering process.

2.4 User-friendly Interface for Non-ML experts

Building and deploying ML models are typically prohibitively complex for all who are not explicitly ML experts; hence, user-friendly interfaces are required to extract critical domain knowledge and to facilitate and instruct ML non-experts with model usage. Some research has attempted to shed light on the obstacles non-professionals experience while using ML models and their perception regarding the function of ML methods [41]. User interaction with ML techniques is extremely useful when comprehending the mechanisms between the ML techniques, and plenty of works can be found in this aspect [9]. There are also works in the medical domain that provides ML non-experts with knowledge of the ML pipeline so that doctors can learn to adopt ML methods for various purposes, including disease detection and diagnosis [39]. Besides ML models, NLP techniques are also vital for health experts since physicians' notes and health records are all unstructured data requiring information extraction. Still, few existing works have tried to bring linguistic methods to non-experts [12]. Clustering or classification frequently follows the extraction of information, which falls in the realm of ML. However, few existing works have integrated NLP and ML

methods into one system that permits direct user-algorithm interaction. Since text data is often the original input in health fields, studies that combine NLP algorithms with ML approaches in its pipeline aimed at disease prediction often exclude the "user-friendly" component and make it hard for physicians or health professionals to actually use those proposed methods [10]. We build a system that is not only accessible to non-experts in NLP and ML by design but also supports the generation and refinement of healthcare indices by incorporating user judgments.

Chapter 3

System Design

The pipeline of ClusT is divided into three stages, the preprocessing phase, the keyword representation phase, and the clustering phase. Prior to the preprocessing stage, we assume the input to the system is a set of geo-tagged tweets that have been retrieved using a pre-defined collection of keywords. For illustrative purposes, we focus on keywords related to the local health environment, including public facilities aimed at enhancing overall health and fitness. The collected Twitter data serves as the input for our system. Figure 3.1 illustrates the overall system pipeline of ClusT.

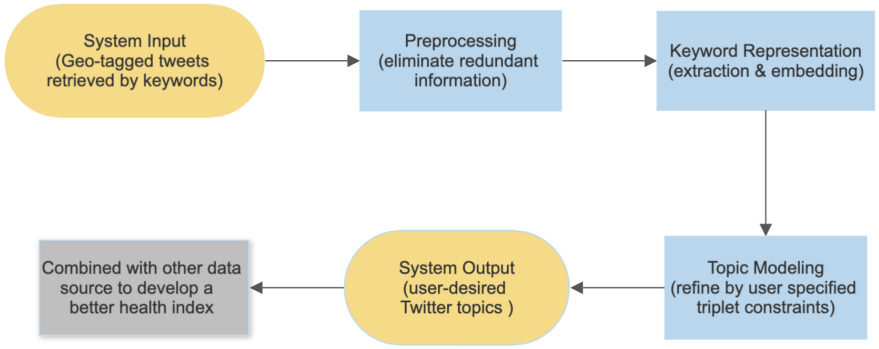

Figure 3.1: System pipeline of ClusT.

3.1 Preprocessing

The preprocessing stage incorporates both user selection and modification capabilities, enabling interactivity and customization. Additionally, user-friendly features will be incorporated to aid individuals who lack expertise in NLP and ML in comprehending the pipeline and manipulating data as desired.

3.1.1 Elimination

Elimination is the process of removing unimportant words from a text so that the output contains more meaningful words. These to-be-removed words can include articles, prepositions, and other useless text components related to the nature of input data [40]. For Twitter specifically, URLs (i.e., www.google.com), hashtags (i.e., #amp), user names (i.e., @emoryuniversity), and emojis are examples of elements that may need to be removed [2].

Figure 3.2 illustrates an example of the elimination process on an original tweet (shown at the top). The process removes two types of elements: (1) the emojis, which are a block of Unicode characters (highlighted in yellow in the figure), and (2) URLs (highlighted in purple). The resulting processed text from removing these special elements is shown at the bottom of the figure.

From a visualization perspective, ClusT provides a tab to customize the elimination steps on the retrieved tweets (illustrated in Figure 3.3). By default, the system removes three types of structural elements: @, hashtags, and URLs. On this page, users can enter regular expressions to eliminate other items (panel B). To clarify the elimination process, the tab displays a random sample of tweets and highlights the words that will be removed (panel C). The highlighted words are color coded to match the structure element or the regular expression. As can be seen from the example, URLs are denoted using purple highlights both in the definition of the structure ele-

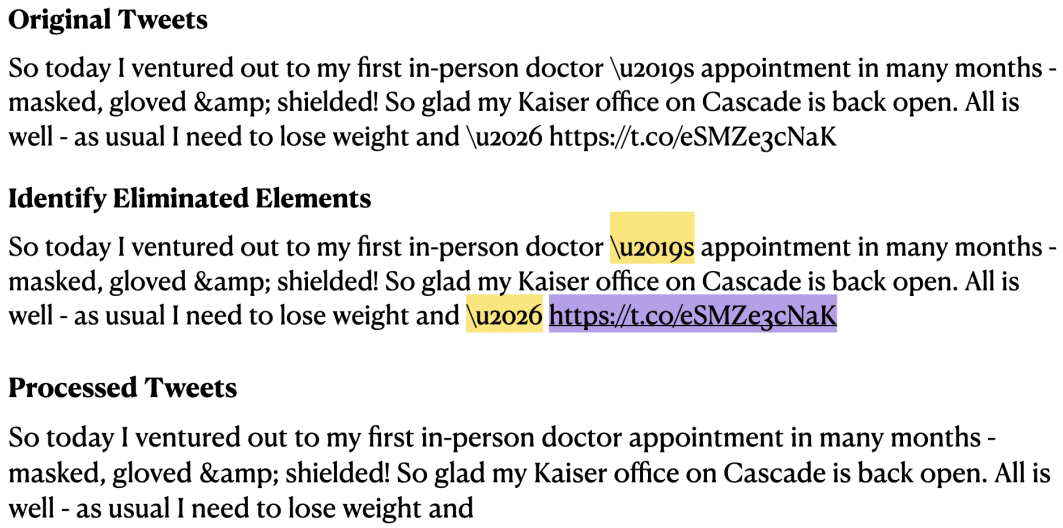

Figure 3.2: An example of the effect of elimination. Two types of elements are removed from the original tweets: (1) emojis, which are blocks of Unicode characters highlighted in yellow, and (2) URLs, highlighted in purple.

ments (panel A) as well as the tweets (panel C). Additional instructions are provided to accommodate the needs of NLP non-experts (at the bottom of panel C).

NLP pipelines typically include a stop-word removal step before the keyword extraction, which removes words such as prepositions, articles, and pronouns [40]. Our pipeline omits the deletion of stop words because our later keyword extraction models often contain this step. After the user is satisfied that all redundant information has been eliminated, they can click the "Save & Next Step," and then the preprocessed data will be sent to the backend for later use. This user-guided preprocessing approach ensures that the data is adequately cleaned and streamlined, paving the way for more accurate and meaningful analysis during the clustering phase.

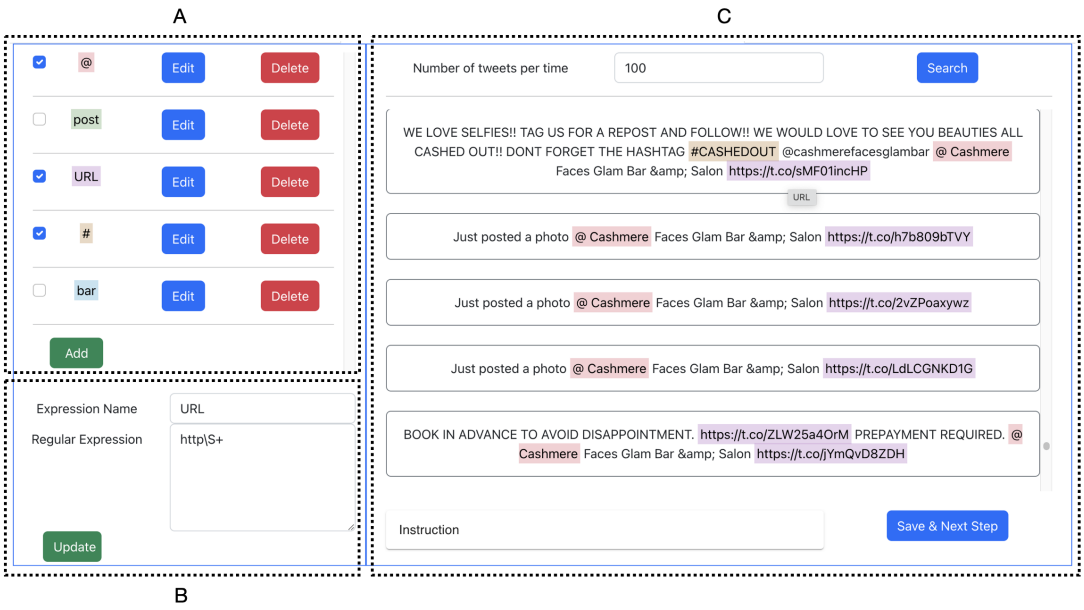

Figure 3.3: Preprocessing tab in ClusT. This tab includes (A) an overview of all elimination elements, (B) a form for adding new elimination elements, and (C) a main view demonstrating the effects of elimination elements on tweets.

3.2 Keyword Representation

3.2.1 Keyword Extraction

Before information extraction, the data consists of many words that can make it difficult to derive meaningful and useful topics. Extraction of keywords or key phrases is an essential step to extract important and relevant terms from each data sample. Various keyword extraction models have been developed to automatically identify the most relevant words and phrases, including the Python Keyphrase Extraction (PKE) module, a versatile Python-based toolkit that accommodates all existing keyphrase extraction methods, both supervised and unsupervised, and the extension to include new approaches [6]. Specifically, we utilize the MultipartiteRank model as it has achieved exceptional performance on numerous prevalent datasets [14] and is highly compatible with Twitter data.

Figure 3.4 provides an example of the results from the keyword extraction pro-

Tweet 1: Can you DM me the name of the hospital? I had neutropenic fever during cancer treatment.

Tweet 2: January 1st ya boy is gone actually have health insurance I haven't seen a doctor in ages.

Figure 3.4: An example of the keywords identified by the MultipartiteRank model. Keywords are highlighted in yellow for each tweets.

cess. The extracted keywords offer a brief overview of the main topics discussed in the tweets and demonstrate a high level of agreement with human judgment. Specifically, keywords such as "hospital," "Neutropenic fever," and "Cancer treatment" are extracted from the first tweet, while keywords like "doctor," "Health insurance," and "ages" are extracted from the second tweet. It should be noted that the keywords extracted by MultipartiteRank for a given tweet may differ from the predefined keyword used to curate that tweet in the sense that they may or may not include the predefined keyword. For instance, the first tweet shown in Figure 3.4 was retrieved using the keyword "hospital," but when keyword extraction was applied to it, more than just the word "hospital" was extracted.

3.2.2 Word Embedding

In order to manipulate the extracted keywords, a numeric representation is needed. A popular method for text representation is word embeddings, which are dense low-dimensional vector representations of words that can preserve semantic characteristics. GloVe (Global Vectors for Word Representation) [27] is a commonly used unsupervised learning algorithm for generating such embeddings. It incorporates co-occurrence statistics in a unique way to produce high-quality word embeddings that capture both syntactic and semantic information. Nonetheless, GloVe is limited to generating embeddings for individual words, which poses a challenge when attempt-

ing to represent multi-word keyphrases. To address this limitation, we compute the average embedding for the words within a keyphrase to derive a final keyphrase embedding. This approach provides a straightforward and effective means to extend the benefits of GloVe to multi-word expressions, ensuring that keyphrases are also represented in a semantically and syntactically meaningful manner within the high-dimensional space.

3.3 Clustering

Given the extracted keywords and their embeddings from each tweet, the goal is then to identify common topics (i.e., clusters) that appear amongst the data. These extract topics can then provide more fine-grained information about the neighborhood environment that can be used to personalize treatments for patients with HF.

3.3.1 Auto-encoder

One limitation of directly using the GloVe embedding is that it is not easily extensible for user input, and there can be a disconnect between the system visualization and the actual word embedding (which is higher than 3 dimensions). Thus, we use an autoencoder to further compress the GloVe embedding. An autoencoder consists of an encoder and a decoder, which work together to compress the input data into a lower-dimensional representation and then reconstruct the original data from this compressed representation [3]. The traditional autoencoder learns the compression by minimizing the reconstruction error between the original input data and the reconstructed output data. Formally, let X denote the input vector. The autoencoder function (i.e., neural network parameters), f, is then trained using the following objective:

$$\mathcal{L} = \min ||X - f(X)||_1 \tag{3.1}$$

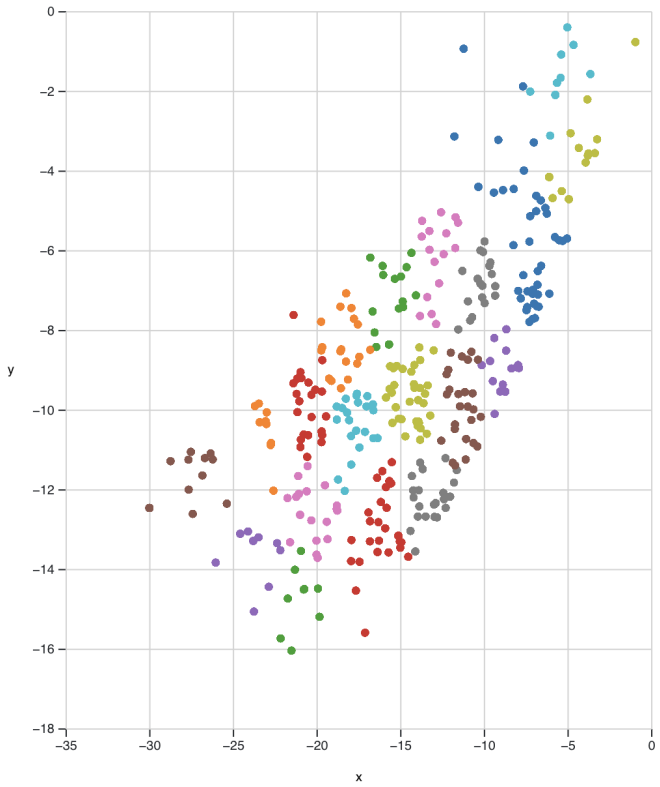

Figure 3.5: This figure illustrates an example of the initial visualization produced by the autoencoder and clustering algorithm. Each point on the plot represents a keyword and is color-coded according to the cluster (or topic) to which it belongs.

By doing so, the autoencoder learns a compressed representation of the input data that captures the most important features of the data while discarding the noise and redundancy.

Prior to user input, we train the auto-encoder model using the Mean Absolute Error (MAE), which is shown in Equation (3.1) and implement early termination to avoid overfitting [25]. The dimension of the compressed embedding is set to two, enabling us to plot keywords in a two-dimensional scatterplot, which serves as the initial visualization for users to explore. Figure 3.5 provides an example of the keyword representations on a two-dimensional scatterplot.

3.3.2 Clustering Algorithm

Once the word embedding has been shrunk to two dimensions, we can apply clustering algorithms to group similar words into coherent topics. Clustering algorithms can be used to partition data points into groups based on their similarity, making it easier to identify patterns and trends within the data.

To effectively capture topic heterogeneity and accommodate the diverse nature of topics within the data, we employ affinity propagation using Euclidean distance on the compressed embeddings. Affinity propagation is a clustering algorithm that identifies representative points, known as exemplars, for each cluster without requiring the number of clusters to be predefined [13]. By applying this technique, we can automatically discover the optimal number of clusters that best represent the underlying topics present in the Twitter data.

Panel A in Figure 3.6 displays all topics, the number of keywords each topic contains, and every individual keyword, allowing users to quickly skim through without being overwhelmed by the density of points in the scatterplot. To swiftly identify keywords of interest, users can either click on the individual keyword in panel A or use the search box at the top of panel B. The scatterplot will then center and zoom in on the selected keyword.

3.3.3 Interactive Keyword Clustering

The clustering results may not be ideal, as some words might not be well-suited for the specific topic. Thus, the idea is to have the user provide input to ClusT to help guide the clustering process. From the machine learning perspective, we adopt the contrastive learning approach to push similar keywords together while also distinguishing the representations between dissimilar keywords. Contrastive learning is a type of unsupervised learning that involves training a neural network to distinguish between similar and dissimilar pairs of data samples. This approach has gained signif-

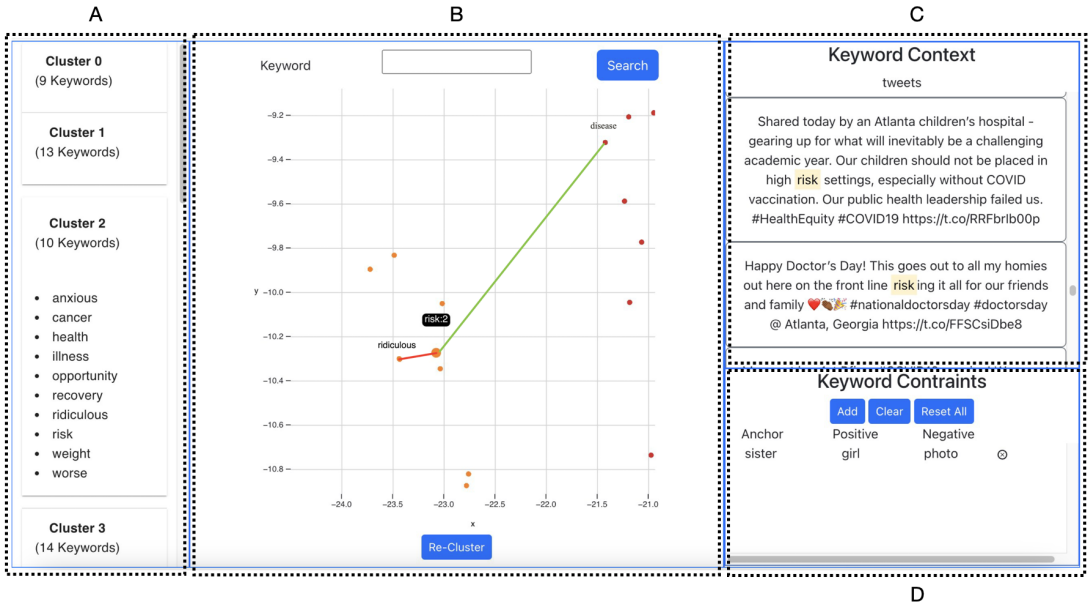

Figure 3.6: Interactive clustering tab in ClusT. The tab consists of four components: (A) a view that displays all topics and keywords, (B) a main view that allows users to explore clusters and add constraints, (C) contextual tweets that contain the selected keywords, and (D) a list of existing constraints that can be manipulated as needed.

icant popularity in the field of computer vision and has been applied to a wide range of tasks such as image classification, object detection, and semantic segmentation.

Triplet loss, introduced by Schroff *et al.* [30], is a popular loss function that was recently shown to be a special case of contrastive loss [18]. Given a reference point (i.e., anchor) A, a positive or matching input P, and a negative or non-matching input N, the goal is to minimize the distance between the anchor and positive point (i.e., $||A - P||_2$), while maximizing the distance between the anchor and negative point (i.e., $||A - N||_2$). Formally, this often takes the form:

$$\mathcal{L} = \max(||f(A) - f(P)||_2 - ||f(A) - f(N)|| + m, 0), \tag{3.2}$$

where m defines the margin or minimum distance between the positive and negative pairs.

To integrate user input into our cluster model, ClusT expects the user to select an

anchor, positive, and negative keyword. Users can select three points (or keywords) in the anchor, positive, and negative order, and then click the 'add' button in panel D to include the triplet constraint. The user-specified triple implies that the anchor and positive are keywords that should be closer together (e.g., within the same cluster), whereas the anchor and negative are keywords that should be further apart (e.g., in separate clusters). To make the constraints clear, a green line will connect the anchor and positive point, while a red line will connect the anchor and negative point. Figure 3.6 illustrates the example of a triplet between risk, disease, and ridiculous.

Users are encouraged to provide as many triplet constraints as they desire, enabling the system to learn a more refined topic model. Users also have the option to start over or remove individual constraints using buttons in panel D. Panel C displays contextual information from tweets containing the selected keywords, assisting users in determining the constraints they wish to add. Once they feel that an adequate number of constraints have been added based on the current plot, users can click the "re-cluster" button (at the bottom of panel B).

The re-cluster button re-trains the autoencoder using the triplet loss, or Equation (3.2). The initial parameters of the autoencoder are saved from the previous training (e.g., MAE loss or other triplet loss). In this manner, the learned representations of the keywords for anchor and positive samples will be pushed together in the lower dimensional space while the anchor and negative samples will be further apart. When the clustering algorithm is re-run on the new keyword embeddings, it is more likely the anchor and positive samples will belong to the same cluster as their distances will be smaller. Users can continue to add constraints and re-cluster until they believe the resulting clustering aligns with their perception of meaningful and consistent groupings. By consistently seeking user input, our model can be refined to generate topics that better align with user preferences.

Chapter 4

User Study

This chapter outlines the user study plan and includes a hypothetical evaluation of the system.

4.1 Study Design

In order to evaluate the validity and usability of ClusT, we plan to conduct a user study that involves participants from diverse backgrounds. Our recruitment efforts will target both machine learning experts from the Computer Science Department of Emory University and healthcare professionals (non-ML experts) from Emory Hospital.

The study will take place in our lab, where participants will begin by viewing a tutorial on how to use ClusT, along with a brief introduction that explains the system's purpose and significance. After the tutorial, participants will be given thirty minutes to explore the system and familiarize themselves with its functionalities and features.

Throughout this exploration phase, a researcher from our team will be available in the lab to address any questions or concerns that participants may have. Upon completing the exploration, participants will be given an additional hour to use ClusT

to train a topic model that aligns with their beliefs and understanding of the data.

Finally, participants will be asked to complete a brief survey that covers their technical background, their experience with the system's usability, the reasonableness of the final topics they generated, and any other aspects that can help us evaluate the effectiveness of ClusT.

4.2 Evaluation Plan

4.2.1 Validity of ClusT

During the user study, each participant will train a topic model. This model will then be applied to Census-based tweets to derive topics. Our approach is based on the methodology used in our preliminary research [42], where the resulting Twitter topics will be combined with other data sources to form a new health index. The performance of this new index will be compared with the traditional Census-based measure of population deprivation, SDI, in terms of predicting 30-day readmission following hospitalization with HF.

To evaluate the validity of our new index, both the ClusT index and the SDI will be fitted into a logistic regression model. Correlations, Area Under the Receiver Operating Characteristic Curve (AUC) values, and R-squared values will be used as evaluation metrics to help us assess the validity of the ClusT index. Given that each participant will train a model and thus generate a new index, we will gather all the evaluation metrics for those indexes to run hypothesis testing. The null hypotheses will be (1) that there is no correlation between the ClusT index and the 30-day readmission following hospitalization with HF, and (2) ClusT index does not have better performance compared to the traditional measure SDI.

By combining these diverse data points, we aim to generate a new health index that can yield more accurate predictions of HF patients' outcomes.

4.2.2 Usability of ClusT

To evaluate the usability of the ClusT system, especially for non-ML experts, we will employ a combination of quantitative and qualitative methods, focusing on the following aspects:

- Ease of learning: Assess how quickly and easily non-ML experts are able to grasp the functionalities of the system by monitoring their progress during the tutorial and exploration phases. Track the number of questions or requests for clarification they have to determine the clarity of the instructions provided.

- Efficiency and effectiveness: Measure the time taken by non-ML experts to complete tasks using the system, as well as the quality of the resulting topic models based on human judgment, perplexity, and coherence. Compare these results with those of ML experts to identify potential gaps in understanding or areas where the system may need improvement to better cater to non-experts.

- Survey responses: Carefully examine the responses from non-ML experts in the post-study survey, focusing on aspects related to the usability of the system, the reasonableness of the generated topics, and any challenges they faced while using ClusT.

By analyzing these aspects, we can evaluate the usability of ClusT and identify areas for improvement. This will allow us to refine the system and make it more accessible and user-friendly for individuals without a strong background in machine learning, and thus achieves its intended goals of generating meaningful, neighborhood-reflective healthcare indices.

Chapter 5

Discussion

Departure from Twitter A huge amount of Twitter users, particularly academics and scientists, are reportedly switching to another social media named Mastodon after Elon Musk took over Twitter in October 2022 [21]. The majority of them made this decision due to two main concerns. For one, Musk has identified himself as a free speech absolutist, so many people are worried that Twitter may further deteriorate with more instances of hate speech and online harassment [21]. For another, many researchers believe that misinformation will become a huge issue if experts are all moving their social media accounts away from Twitter, and the platform will be flooded with irrational voices. If Twitter experiences a significant loss of academic users, the data we retrieve will likely be skewed towards people with weaker academic backgrounds, thus introducing a bias. The authenticity of the information provided by tweets will also be impacted negatively. Twitter may also experience a rise in tweets, including misspelled terms, implying that we may encounter issues when performing keyword extractions. If Twitter's number of active users continues to decline, the size of our usable data will likely shrink.

Bias in Geotagged tweets Accessing the sent location of some tweets was one of the key prerequisites for this study. However, a previous study has shown that

approximately only 0.85% of tweets are geotagged [33], which means our data pool is constrained. Furthermore, users of geotagged tweets are not randomly distributed across the US population [23]. Another study has shown that age, race, income, and area of living are all factors associated with the likelihood for a user to make their tweets geotagged. These correlations suggest that the tweets we retrieved will be skewed towards specific social groups, affecting the ultimate healthcare index. For instance, it was found that people of color, including Asians, Hispanics, and Blacks, are more likely to enable geotagged, suggesting that the healthcare index we construct may be more reflective of the health resources accessible to these groups of people [23].

Tweets' Retrieval Because we retrieve tweets using public facilities as keywords, only tweets containing the name of the facility's category will be extracted. A common form of a tweet that we get is, therefore, "I went to a restaurant yesterday." This, however, implies that we have excluded tweets that directly include the facility's name, such as McDonald's, as an example of a restaurant. One may imagine public-health-related keywords to be mentioned by sentences like "I'm going to the park." but in reality, it is uncommon for people to use "the park" to refer to a specific park on social media sites. Additionally, even if words like hospital and park are included in the facility's name, it is still likely for Twitter users to type an abbreviation due to the language used on Twitter often being relatively colloquial. As a result of these complexities, not only do we fail to acquire tweets that may contain valuable information, but our data size also decreases.

Tweets' Content Our study aimed to use the content of tweets to examine the state of health infrastructure in diverse neighborhoods. However, upon analyzing the data, we found that the majority of tweets were focused on personal health issues rather than providing insights into the local healthcare system. Specifically, we ob-

served that many tweets expressed concerns such as "I haven't seen a doctor in ages," which suggests that access to healthcare services may be limited or inadequate for certain individuals. Although social media data can be a valuable resource for public health research, our findings indicate the need for further investigation to distinguish between tweets related to health infrastructure and those related to personal health concerns.

General Data Quality During the developmental stage of the system, we noticed that many of the random tweets shown in the preprocessing tab are duplicates, the majority of which are bot tweets that are either advertisements or notifications saying a photo has been posted. Tweets like this are incapable of delivering any relevant information, so we deemed it necessary to remove them after the elimination stage. Because elements such as user names and URLs are already removed in the elimination phase, bot tweets are likely to be exact duplicates of one another, making it easier for us to remove them simultaneously. Even during the preprocessing phase, we may need some NLP algorithms to display tweets that are different from each other such that users can gain a more comprehensive view of the data.

Future Work Our system employs NLP and ML approaches to provide public health professionals without CS expertise access to constructing and refining a more localized health index via visualizations. To satisfy the needs of non-ML experts, we designed all system components with user-friendly characteristics and established straightforward interaction methods. However, our work has certain limitations that we intend to address in future studies.

The first shortcoming is the inadequate sample size. Using census-based geotagged tweets as our data already implies a relatively small data size, and retrieving only those containing public facilities worsens the situation. We noticed that it is at times likely for tweets retrieved using certain keywords for a given census tract to be non-

existent, making it hard for keyword extraction models to be performed. In the future, we aim to design a more detailed data retrieval method such that we have more data for constructing a healthcare index.

The second issue is the nature of the social media data—with users possessing certain demographic characteristics. Unlike census data, from which SDI is derived, Twitter users are not distributed uniformly across all age groups and socioeconomic classes. Since our study is concerned with the healthcare index, which is directly correlated with the socioeconomic status of the community users are from, having data that is skewed toward people with certain demographic features implies that the healthcare index we derive is no longer a measure of the community's access to health resources but that of certain groups. To overcome this issue, we may want to combine SDI or other existing measures with social media data in the healthcare index's generation to ensure that all demographic groups are represented in some way.

The third issue we encountered was the complexity of interpreting tweets. For instance, the tweet "January 1st ya boy is gone actually have health insurance. I haven't seen a doctor in ages" not only indicates a change in the individual's insurance status but also suggests a long gap in receiving medical attention. This nuanced information cannot be effectively retrieved through keywords such as "doctor" or "health insurance." To address this challenge, we plan to incorporate sentiment analysis on keywords or retrieve more contextual embedding through BERT models.

Based on the feedback received during our formative evaluation, we are taking steps to enhance our system. One such improvement involves enabling users to input either anchor and positive keywords or anchor and negative keywords and to remove specific keywords during the topic modeling phase. As our system uses soft constraints, there is a possibility that some user inputs may not be fully satisfied. To mitigate this, we plan to implement a feature that enables users to prioritize certain constraints or receive warnings if certain constraints cannot be met.

Moreover, we aim to extend our system's application to other text media such as Facebook and Reddit to incorporate more information into the health index generation process.

Chapter 6

Conclusion

Accurate assessment of the neighborhood environment not only provides a way for public health experts to understand regional accessibility to essential health resources but also identifies correlations between the environment and the likelihood of relapse and readmission following treatment. In addition, more effective treatment strategies can be devised with this knowledge to help patients get the customized interventions necessary for recovery. We designed a visual analytic system, ClusT, that can be employed in developing and refining a neighborhood environment index based on Twitter data. To enable participation from health experts without a computer science background and to leverage their unique perspectives, we have implemented visualizations to facilitate user customizations of natural language processing and machine learning algorithms. Our approach ensures that diverse insights can be integrated into the analysis, leading to a more comprehensive understanding of the factors influencing healthcare outcomes. We believe this innovative approach has the potential to revolutionize healthcare research and contribute to more informed decision-making in public health policy and treatment strategies.

Bibliography

[1] Elvis A Akwo, Edmond K Kabagambe, Frank E Harrell Jr, William J Blot, Justin M Bachmann, Thomas J Wang, Deepak K Gupta, and Loren Lipworth. Neighborhood deprivation predicts heart failure risk in a low-income population of blacks and whites in the southeastern united states. *Circulation: Cardiovascular Quality and Outcomes*, 11(1):e004052, 2018.

[2] Giulio Angiani, Laura Ferrari, Tomaso Fontanini, Paolo Fornacciari, Eleonora Iotti, Federico Magliani, and Stefano Manicardi. A comparison between preprocessing techniques for sentiment analysis in twitter. In *KDWeb*, 2016.

[3] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders, 2021.

[4] Lily S Barnard, Deborah J Wexler, Darren DeWalt, and Seth A Berkowitz. Material need support interventions for diabetes prevention and control: a systematic review. *Current diabetes reports*, 15:1–8, 2015.

[5] Sugato Basu, Arindam Banerjee, and Raymond J. Mooney. Active semi-supervision for pairwise constrained clustering. *Proceedings of the 2004 SIAM International Conference on Data Mining*, 2004.

[6] Florian Boudin. pke: an open source python-based keyphrase extraction toolkit. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 69–73, Osaka, Japan, December 2016.

[7] US Census Bureau. American community survey (acs), Mar 2023.

[8] Danielle C Butler, Stephen Petterson, Robert L Phillips, and Andrew W Bazemore. Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery. *Health services research*, 48(2pt1):539–559, 2013.

[9] Duen Horng Chau, Aniket Kittur, Jason I Hong, and Christos Faloutsos. Apolo: making sense of large network data by combining rich user interaction and machine learning. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 167–176, 2011.

[10] Pratiksha R Deshmukh and Rashmi Phalnikar. Information extraction for prognostic stage prediction from breast cancer medical records using nlp and ml. *Medical & Biological Engineering & Computing*, 59(9):1751–1772, 2021.

[11] Ana V. Diez Roux and Christina Mair. Neighborhoods and health. *Annals of the New York Academy of Sciences*, 1186(1):125–145, 2010.

[12] Gail Dutton. Big pharma reads big data, sees big picture: Linguamatics brings natural language processing to non-experts, expediting drug development. *Genetic Engineering & Biotechnology News*, 38(1):8–9, 2018.

[13] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[14] Ygor Gallina, Florian Boudin, and Béatrice Daille. Large-scale evaluation of keyphrase extraction models. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 2020.

[15] Louis Hickman, Stuti Thapa, Louis Tay, Mengyang Cao, and Padmini Srini-

vasan. Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 25(1):114–146, 2022.

[16] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. Interactive topic modeling. *Machine Learning*, 95(3):423–469, 2013.

[17] Heval M Kelli, Muhammad Hammadah, Hina Ahmed, Yi-An Ko, Matthew Topel, Ayman Samman-Tahhan, Mossab Awad, Keyur Patel, Kareem Mohammed, Laurence S Sperling, et al. Association between living in food deserts and cardiovascular risk. *Circulation: Cardiovascular Quality and Outcomes*, 10(9):e003532, 2017.

[18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[19] Hannah Kim, Dongjin Choi, Barry Drake, Alex Endert, and Haesun Park. Topicsifter: Interactive search space reduction through targeted topic modeling. *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2019.

[20] Minjeong Kim, Kyeongpil Kang, Deokgun Park, Jaegul Choo, and Niklas Elmqvist. Topiclens: Efficient multi-level visual topic exploration of large-scale document collections. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):151–160, 2017.

[21] Kai Kupferschmidt. As musk reshapes twitter, academics ponder taking flight. *Science (New York, NY)*, 378(6620):583–584, 2022.

[22] Nicole I Larson, Mary T Story, and Melissa C Nelson. Neighborhood environments: disparities in access to healthy foods in the us. *American journal of preventive medicine*, 36(1):74–81, 2009.

[23] Momin Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. Population bias in geotagged tweets. In *proceedings of the international AAAI conference on web and social media*, volume 9, pages 18–27, 2015.

[24] Lynne C Messer, Barbara A Laraia, Jay S Kaufman, Janet Eyster, Claudia Holzman, Jennifer Culhane, Irma Elo, Jessica G Burke, and Patricia O'campo. The development of a standardized neighborhood deprivation index. *Journal of Urban Health*, 83(6):1041–1062, 2006.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[26] Shivani A Patel, Maya Krasnow, Kaitlyn Long, Theresa Shirey, Neal Dickert, and Alanna A Morris. Excess 30-day heart failure readmissions and mortality in black patients increases with neighborhood deprivation. *Circulation: Heart Failure*, 13(12):e007947, 2020.

[27] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[28] Margaret Quinn Rosenzweig, Andrew D. Althouse, Lindsay Sabik, Robert Arnold, Edward Chu, Thomas J. Smith, Kenneth Smith, Douglas White, and Yael Schenker. The association between area deprivation index and patient-

reported outcomes in patients with advanced cancer. *Health Equity*, 5(1):8–16, 2021.

[29] AV Diez Roux and Christina Mair. Neighborhoods and health. *Ann NY Acad Sci*, 1186(1):125–145, 2010.

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[31] Lisa B Signorello, Sarah S Cohen, David R Williams, Heather M Munro, Margaret K Hargreaves, and William J Blot. Socioeconomic status, race, and mortality: a prospective cohort study. *American journal of public health*, 104(12):e98–e107, 2014.

[32] Gopal K. Singh. Area deprivation and widening inequalities in us mortality, 1969–1998. *American Journal of Public Health*, 93(7):1137–1143, 2003.

[33] Luke Sloan, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological research online*, 18(3):74–84, 2013.

[34] Brittany L Smalls, Chris M Gregory, James S Zoller, and Leonard E Egede. Effect of neighborhood factors on diabetes self-care behaviors in adults with type 2 diabetes. *Diabetes research and clinical practice*, 106(3):435–442, 2014.

[35] Brittany L Smalls, Chris M Gregory, James S Zoller, and Leonard E Egede. Assessing the relationship between neighborhood factors and diabetes related health outcomes and self-care behaviors. *BMC health services research*, 15(1):1–11, 2015.

[36] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. Closing the loop. *23rd International Conference on Intelligent User Interfaces*, 2018.

[37] Joni L Strom and Leonard E Egede. The impact of social support on outcomes in adult patients with type 2 diabetes: a systematic review. *Current diabetes reports*, 12:769–781, 2012.

[38] M Sukanya and S Biruntha. Techniques on text mining. In *2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, pages 269–271. IEEE, 2012.

[39] Andrea Vázquez-Ingelmo, Julia Alonso-Sánchez, Alicia García-Holgado, Francisco José García Peñalvo, Jesús Sampedro-Gómez, Antonio Sánchez-Puente, Víctor Vicente-Palacios, P Ignacio Dorado-Díaz, and Pedro L Sanchez. Bringing machine learning closer to non-experts: proposal of a user-friendly machine learning tool in the healthcare domain. In *Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)*, pages 324–329, 2021.

[40] S Vijayarani, Ms J Ilamathi, Ms Nithya, et al. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2015.

[41] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 designing interactive systems conference*, pages 573–584, 2018.

[42] Juhao Zhang, Samantha Lin, Yunjie Wu, Jing Zhang, Alanna Morris, Shivani Patel, and Joyce Ho. Abstract 15011: Deriving and validating novel neighbor-

hood data for investigation of adverse outcomes in patients hospitalized for heart failure: A feasibility study. *2022 AHA*, 2022.