INST327, Section: 0104
Final Project Report
12/09/2024
Team 3: Tanushree Sharma, Ramya Nataraj, Ayana Khan, Gabrielle Waugh, Myles Sartor

Introduction:

Americans apart of minority demographics all over the nation have consistently had to grapple with the harsh realities of their situation when it comes to the level of targeted hate toward them in their respective communities. Whether it is reported or not, hate crimes have been an unfortunate facet of the United States for quite some time now. As individuals within minority groups ourselves, we have been physically, mentally, and emotionally influenced by the presence of these hate crimes, and we recognize the importance of taming such behaviors internally to respect the diversity of our country. Therefore, we settled on this topic, which was quite relevant to the frightening amount of hateful events still being presented in our news in 2024. Our database was designed to document and analyze certain wrongdoings while simultaneously providing the means to measure the impact of such violent episodes over time. In creating this compound dataset full of granular complexity, we will not only be creating a more safeguarded environment for minority groups by shedding light on these hate disparities, but we'll also be helping to uncover the harsh truths that exist for the particular groups that are discriminated against so that all Americans can become aware of the injustices currently taking place in the United States. After all, the data that provides the general facts about the frequency in which these events take place will be pivotal in understanding the underlying nature of them.

The original dataset (ProPublica) includes hate incidents that took place from February 13th, 2017, to August 12th, 2017, giving us an expansive view of the totality of discrimination that took place in that given period. It holds various instances of crimes reported by organizations against a minority population. The large amount of reported incidents available for analysis allows us to identify verifiable trends or patterns that exist in the original dataset.

The information within our database is formatted in such a way that allows for streamlining and improving the original dataset for practical use by those pushing for civil changes and justice. Highly focused usage of the third normal form and practically integrated changes for the original attributes in the "Documenting Hate" database make information-seeking more pertinent and useful overall for users. Our newly shaped dataset looks to trim some of the obscurity surrounding this topic so that we can effectively tame the severity of the issues at hand while attempting to advocate for change and support victims civilly and legally.

Database Description:

The database we constructed was designed to document and analyze certain wrongdoings while simultaneously providing the means to measure the impact of such violent episodes over time. As a tool for users looking to shed light on the reality of the situation, it holds seven tables with information on 45 articles and organizations from 35 different locations in the United States. From this, one can get a good understanding of how many types of different hate crimes are being committed against specific groups in certain cities.

Logical Design:

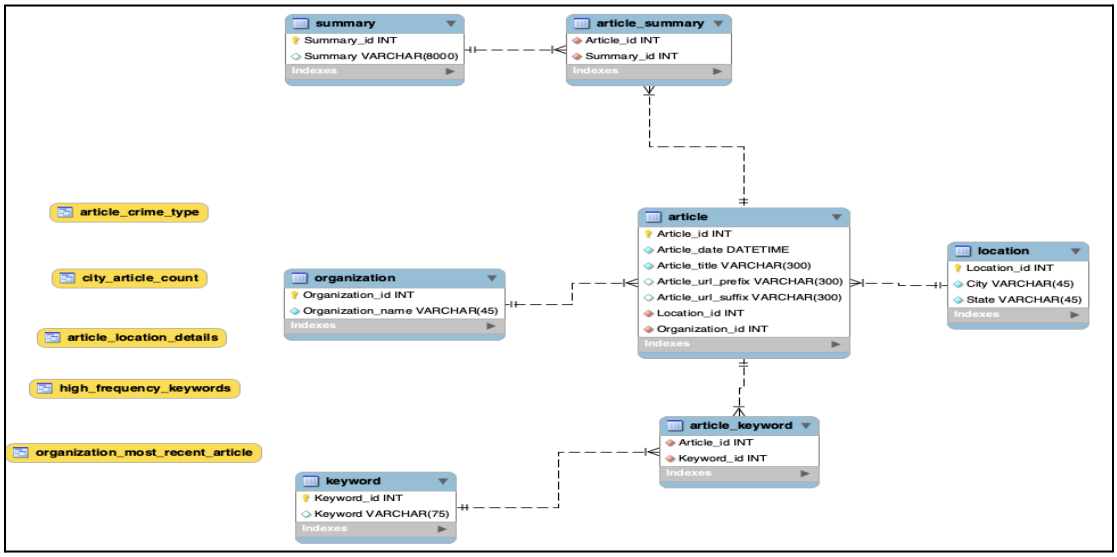

Image 1. Entity Relationship Database for Documenting Hate Database.

We are organizing this database to allow people to identify various city hotspots with hate crimes. By breaking down this database, we separate articles from their organization and locations to gain a more nuanced understanding of the information provided.

Our team decided to give users a broad enough scope of the represented data while understanding the relationship between each table. The importance of some foreign and primary keys (auto-incremented) was evident through analyzing the various data and writing the queries. 'Article_id' was utilized as a primary key in its table and as a foreign key in two linking tables, 'Article_Summary' and 'Article_Keyword,' which defines the many-to-many relationship between articles to summaries and articles to keywords. Furthermore, 'Location_id' and 'Organization_id' were further created as foreign keys in the Article table to create a one-to-many relationship for organization-to-article and location-to-article. These were to satisfy

discussions on where, when, and by whom each article was written. While creating the database, it was also decided that all IDs, Locations, and Organizations would be created as "not null," allowing us to classify which articles we had enough information for.

By detailing this process extensively with the proper keys and connections between tables using certain relationships, it became easier for database users to quickly query whatever information they might've wanted without having to write an overwhelming amount of different subqueries. For instance, a query fueling a user's desire to find the number of hate crimes involving racists stabbing someone in New York could be performed smoothly with no functionality impeded in the process.

Physical Database:

Although our final database consists of seven tables: article, organization, location, keyword, summary, article_summary, and article_keyword, there were no one-to-one relationships found in the database. With all of them as one-many or many-many relationships, the data management (Less redundancy, more reusability), organization, scalability, and flexibility were much better. We split the URL attribute into its prefix and suffix parts and related everything else that had to do with the article itself back to article_id within the article table. All the attributes that depended on location_id and organization_id were placed within their own location and organization tables to allow for more meaningful relationships with the article table. For instance, this is extremely necessary if someone is interested in uncovering the truth about hate crimes in a particular location or organization. In covering the potential possibilities for the use of this data, it was imperative for us to narrow down what we incorporated into the database while still maintaining practicality. In addition, summary and keyword were given their own tables to account for situations where certain summaries and keywords are reused. The linking tables article_summary and article_keyword would then reduce this data redundancy further by giving summaries and keywords their own specific columns. Whenever there are multiple summaries and keywords for a single article or multiple articles with the same summary and keywords utilized, this setup helps keep track of that process. It validates that data as it is scaled over time.

Sample Data:

The data for hate crimes we included in our database is from ProPublica, a nonprofit organization. These are crimes committed between March 24th, 2017, and March 27th, 2017,

and we narrowed the dataset to various articles with information for all attributes: article date, title, organization, city, state, URL prefix/suffix, keywords, and the summary.

We improved our original Documenting Hate database by creating an environment of the third normal form tables in our current ERD model. We created an article table containing all relevant information for the articles in the database. The foreign id keys link the article table to the locations and organizations table, and separate views of both tables provide the information pertaining to them. Other tables, like keyword and summary make good use of their linking tables and display relevant information similarly. The rest of the information directly pertaining to the article was included in the article table, such as the date and time it was released, the title, and the URL components. The separation of components within specific entities kept our dataset from becoming too cluttered and difficult to write queries for in the future (Information gathering, UPDATES/DELETES, and INSERTS).

Location ID	City	State
1	New York	NY
2	Jackson	MS
3	Seattle	WA
4	West Palm Beach	FL

Image 2. A snippet of the location table city-state combinations in which articles were written.

Views and Queries:

Views	Req. A (X4)	Req B. (X3)	Req. C (X2)	Req D. (X1)	Req E. (X1)
high_Frequency_Keywords	X	X	X		
city_article_count	X		X		
organization_most_recent_article	X				
article_crime_type	X	X		X	
article_location_details	X	X	X		X
TOTAL	**5**	**3**	**3**	**1**	**1**

Explanations of All Queries:

Query 1: Creates a view that shows the highest frequency keywords amongst the articles, sorted by the most occurred to least.

Query 2: Creates a view that shows the number of times each location or city-state combo comes up in the list of articles.

Query 3: Creates a view that is ordered by the most recent article written by the organization.

Query 4: Creates a view that shows the summaries and articles related to specific types of crimes referred to in the following keywords: stabbing, murder, attack, and threats.

Query 5: Creates a view that finds details of articles from locations that are associated with more than one article.

Changes From the Original Design:

Our original data from the proposal was impractical, as it included attributes and entities that weren't even a part of the original database (Zip codes, emails, phone numbers, crime type, targeted reasons, etc.). We got rid of most of them and started from the beginning with what we were given, as it would've been impossible for us to have the information that we were thinking of, so we had to change our entire approach to use only what we had been given, while still aligning with project requirements. We already had a small dataset to work with, so altering the original to work with seven entities that process the sample data well enough through linking tables was a major achievement that we have been able to adjust for in our project scope. In making the changes to the nature of our data attributes and consolidating the information we planned on retrieving from the original, we were able to honor what we originally wanted to uncover by studying the data's possible trends while simultaneously boosting its overall utility.

Despite our changes, we would still be able to uncover how many crimes were going on in what specific areas and who was ultimately reporting on them (Organizations). The feedback we got from various TAs and AMPs helped us to take many things into consideration when designing our finalized dataset. They pointed out that it was okay to have some tables with only one or two attributes for our data due to their limited nature. It was more important for us to make the database practical for use and design columns that made joins, updates, and potential deletes much easier to execute. Utilizing advice from one another and advice from the instructional team allowed us to realize these shortcomings, as it helped us focus and ask better questions about our data. If we want to know how many times a certain keyword appears across data or identify an end in summaries for a certain time frame, it is all easily doable now with these new tables.

Diversity, Equity, and Inclusion Considerations:

Our database continues to have more than enough potential to represent a broad spectrum of demographic, historical, and social diversity. We have worked to improve our inclusivity by

including articles from many sources in the original dataset. By doing this, the narrative for those most impacted by these crimes can be broadly interpreted, and the bias of groups can be eliminated. Our dataset is based on news sources, social media, public tips, and law enforcement reports. Nevertheless, we can work to identify publications that document the hate experience the best, even though some may be less reported. By doing so, the marginalized communities will be represented fairly since their interactions with the media or government may already be tense. Along with that, we also decided to use data from both urban and rural locations to fairly represent hate crimes that take place across the country.

With the way we have set up our tables, the goal of our database is to strive to record hate crimes that target not only well-known or bigger demographic groups but also smaller or less-represented communities that may also go unnoticed. Just like the original database, we will achieve this objectively but in a more focused, simplified, and streamlined way. This guarantees that the dataset represents the experiences of all groups impacted by hate crimes in as many locations as possible rather than only those more commonly covered by the media.

Data Privacy, Fair Use, Other Ethical Considerations:

Potential data privacy, fair use, and other ethical concerns regarding this hate crime data set are relative to the intended use of the database. As explained by the Brennan Center for Justice, the collection of hate crime data enables policymakers to consider the experiences of victims in policy-making, allowing them to take the appropriate measures to prioritize the safety and security of marginalized communities affected by hate crimes. However, considerations change relative to data usage in cases where highlighting specific data points may better support an organization’s agenda (i.e., raising awareness of a specific type of hate crime or for a specific targeted group). ProPublica collected the data used in this database, along with tips from the public and information from various security agencies, to develop it further. No data privacy concerns, such as legal infringement, are present in the database (ProPublica), as all contents are public knowledge. However, some summaries contain the names of people involved in such crimes (ProPublica Database), which raises some data privacy and ethical concerns regarding personal information. Despite this, the names remain in the database due to public knowledge of these crimes. Additionally, victims' names are often included in crime reports at the discretion of the victim’s family, so it felt fair to assume that including this information did not fall under privacy infringement.

Lessons Learned:

The development of our database project emphasized the importance of clear communication, adaptability, and ethical considerations. At the start, unclear roles and planning created confusion about the requirements for each assignment. To address this, we implemented a to-do list shared in the group chat after each meeting. This assigned clear responsibilities, set deadlines, and ensured everyone stayed informed, even if they could not attend a meeting. The project required us to adapt to the limitations of our dataset, prompting a redesign focused on the actual attributes available. While we initially wanted to include features like zip codes and other details to draw meaning from the dataset, these elements were not present in the data, so we had to work with what we had. We wrote complex queries and overcame technical issues with MySQL, strengthening our SQL and teamwork skills. When some team members struggled to import data into the ERD, we collaborated on Zoom to create each CRUD sample query as a group. Working with sensitive data also requires careful ethical consideration. For example, while we identified keywords for analysis, we avoided terms related to race or ethnicity to minimize bias. Throughout this project, we learned the value of starting early and dedicating sufficient time to testing and troubleshooting, which helped reduce stress and improve results. This experience significantly enhanced our technical expertise and collaborative skills.

Potential Future Work:

Our database holds endless possibilities of future work that it could be used for, especially in its role in bringing positive change to society. It can become an essential tool for journalists, researchers, advocates, legislators, and more who are working to fight against discrimination. By continuously updating and expanding our database, we will aim to incorporate data from a wider range of time periods and include more detailed information. Our database's documentation of hate crimes not only showcases how serious these incidents are but it also allows users to gain a better understanding of the trends and patterns of targeted violence. With the continuation of more recent information being added, it could help with promoting greater awareness of the hate inequalities around the U.S and advocate for minorities such as us.

References

"Documenting Hate News Index | Propublica." Propublica Documenting Hate Project, 2017, projects.propublica.org/hate-news-index/. Accessed 26 Sept. 2024.

German, Michael, and Emmanuel Mauleón. "Fighting Far-Right Violence and Hate Crimes." Brennan Center for Justice, 1 July 2019, www.brennancenter.org/our-work/research-reports/fighting-far-right-violence-and-hate-crimes. Accessed 26 Sept. 2024.

ProPublica. "Documenting Hate." ProPublica, 8 Sept. 2016, projects.propublica.org/graphics/hatecrimes. Accessed 26 Sept. 2024.