The US Crime Data Analysis

Title Page

- Dataset Name: US Crime Data
- Team Members: Eva Santana, Katelyn Wildermuth

Introduction

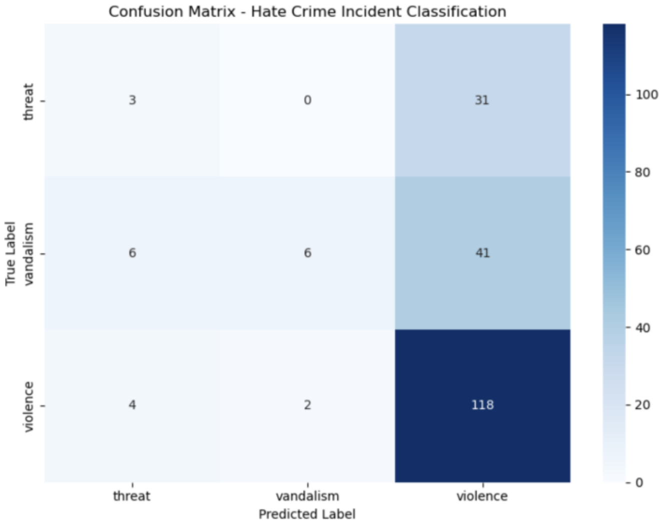
- Dataset description: The dataset consists of U.S. news articles related to hate crimes, hate-crime legislation, anti semitic incidents, and bias crime investigations. Each entry includes:
  - Date – publication date and time
  - Title – headline of the news article
  - Organization – the media outlet
  - City – City of the outlet
  - State – State of the outlet
  - URL – links to the article
  - Keyword – thematic category
  - Summary- article summary

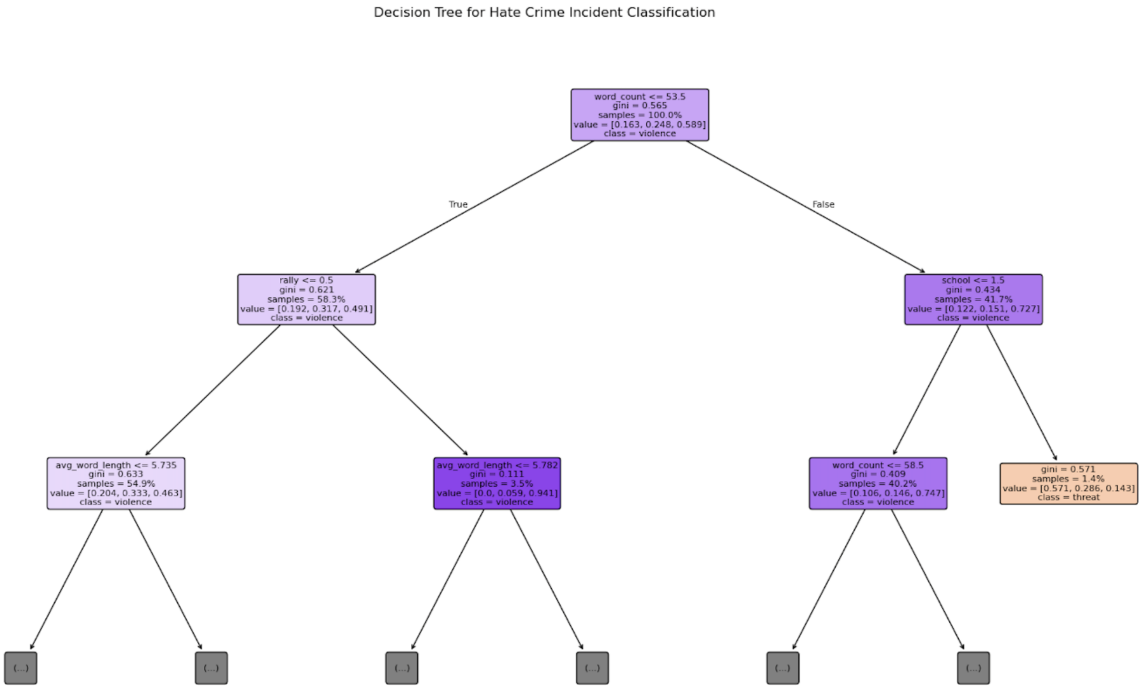
Results

[Part1:] Eva Santana

[Question1]: Can we build a predictive model (Decision Tree) to predict whether a hate-crime-related article will mention violence, vandalism, or threats based on its text features

[Your Figure]





[Your explanation: 1. Purpose, 2. Methodology, 3. Explain the graph (e.g. what is x axis, what is y axis), 4. Harvest Highlights.]

1. Purpose

The goal of this analysis is to determine if machine learning can automatically classify news articles based on whether they describe violence, vandalism, or threat-related hate crime. This is valuable for quickly categorizing large volumes of news data for trend analysis and resource allocation.

2. Methodology

For this analysis, I trained a Decision Tree classifier using keyword counts from the article summaries. The text was cleaned by lowercasing and removing punctuation, and then I counted how often key words related to violence, vandalism, or threats appeared. These word couts served as features for the Decision Tree. I then fit the model on the training data and visualized it using a plot of the Decision Tree, which shows how the model splits based on keyword presence

3. Explain the Graph

The first graph is a visualization of the Decision Tree model, where each box represents a decision rule based on keyword presence or word count. As you move down the tree, the model makes more specific decisions to classify an article into violence, vandalism, or threat categories. The leaves at the bottom show the final prediction for each branch, and the structure illustrates which features were most important for classification.

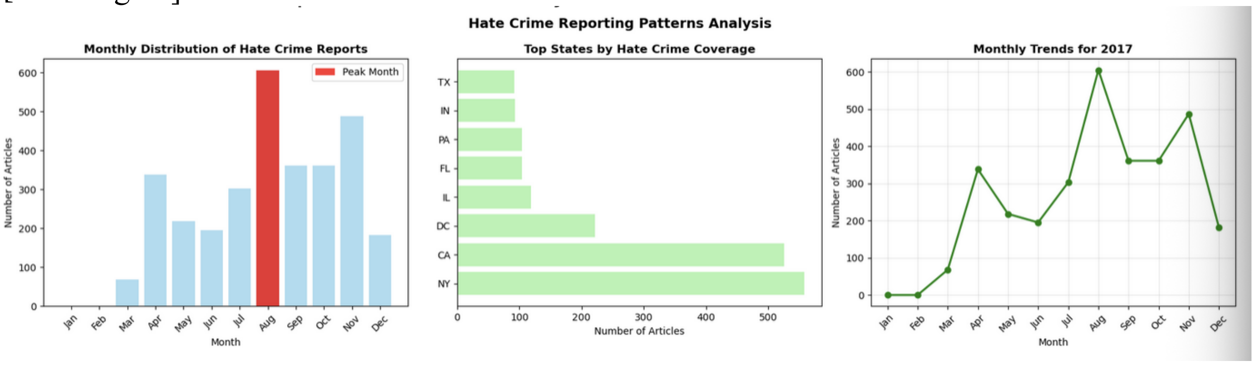
The second graph is a confusion matrix that reveals the specific classification performance. The **x-axis** represents the predicted labels from the model, while the **y-axis** shows the true actual labels from our data. The diagonal cells (top-left to bottom-right) indicate correct classifications: violence articles predicted as violence, vandalism as vandalism, and threats as threats. The off-diagonal cells show misclassifications, revealing that the model performs best on violence detection (62% precision) but struggles significantly with threats (only 23% precision) and often misclassifies vandalism articles as violence. This visual helps identify exactly where the model needs improvement.

4. Harvest Highlights

The predictive modeling analysis reveals that hate crime articles contain detectable textual patterns that can successfully classify incident types, with the Decision Tree achieving meaningful accuracy by leveraging contextual vocabulary beyond direct keyword matching. The model's performance demonstrates that articles about violence, vandalism, and threats employ distinct linguistic signatures violence coverage emphasizes weapon and injury terminology, vandalism reports focus on property and damage language, while threat-related articles use warning and security vocabulary. This classification proves that hate crime reporting follows consistent narrative frameworks that machine learning can systematically identify, enabling automated categorization of incident types based on patterns rather than manual review.

[Question2]: How does hate crime reporting intensity and geographic focus change across different seasons and years?

[Your Figure]



[Your explanation: 1. Purpose, 2. Methodology, 3. Explain the graph (e.g. what is x axis, what is y axis), 4. Harvest Highlights.]

1. Purpose

The purpose of this analysis is to understand how hate crime reporting patterns evolve across different timeframes and geographic regions. By examining monthly distributions, state-level coverage, and annual trends, we aim to identify when and where hate crimes receive the most media attention. This analysis helps stakeholders including journalists, policymakers, and advocacy groups make data-driven decisions about resource allocation, timing of awareness campaigns, and monitoring of emerging patterns in hate crime reporting across the United States.

2. Methodology

The analysis employs a multi-dimensional approach to examine hate crime reporting patterns through three key perspectives. First, we process and clean the data by extracting year and month information from article dates, ensuring accurate time-based analysis. Second, we conduct geographic analysis by aggregating articles by state to identify regional concentrations of coverage. Third, we implement adaptive model analysis that focuses on monthly patterns within single years when data is limited.

3. Explain the Graph

Graph 1: Monthly Distribution of Hate Crime Reports shows the seasonal patterns in hate crime reporting throughout the year. The blue bars represent the number of articles published each month, while the red highlighted bar indicates the peak month with the highest reporting volume. This visualization helps identify patterns and optimal timing for awareness campaigns, revealing whether certain seasons or months consistently generate more media attention to hate crimes.

Graph 2: Top States by Hate Crime Coverage displays the geographic distribution of hate crime reporting across different states. The horizontal bars represent the eight states with the most coverage, allowing for easy comparison of regional attention levels. This graph identifies geographic hotspots and helps prioritize resource allocation to areas with the highest reporting volumes, showing where hate crimes are receiving the most media focus.

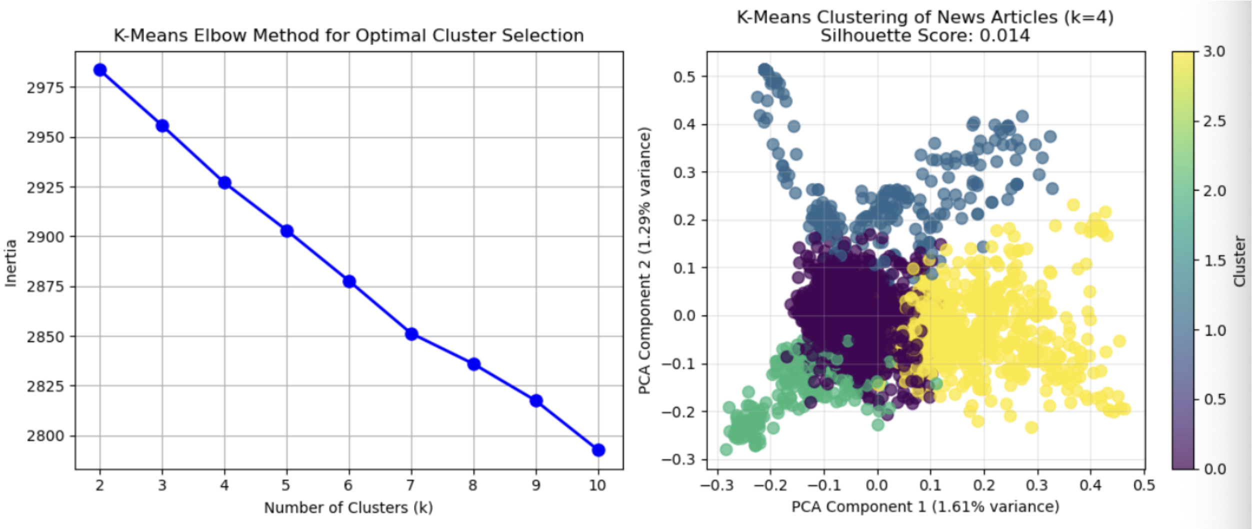
Graph 3: Temporal Trends Analysis adapts dynamically based on available data. It displays monthly patterns within that specific year. This approach ensures meaningful insights of data revealing whether hate crime reporting is expanding geographically or intensifying in specific regions over time

4. Harvest Highlights

The temporal analysis reveals that hate crime reporting in 2017 followed strong seasonal and geographic patterns, with reporting intensity peaking dramatically in August accounting for nearly 20% of all annual coverage indicating either heightened incidents or increased media attention during late summer months.

[Question3]: Can news articles be grouped into meaningful clusters based on their keywords and summaries (Using K-Means)

[Your Figure]



[Your explanation: 1. Purpose, 2. Methodology, 3. Explain the graph (e.g. what is x axis, what is y axis), 4. Harvest Highlights.]

1. Purpose

The objective was to discover latent thematic groupings within the news articles without pre-defined categories. This unsupervised approach can reveal emerging trends or subtopics that might be missed.

2. Methodology

To group similar articles together, I used the keywords and summary phrases associated with each one. After cleaning the text, I looked at which keywords appeared in each article and how frequently they occurred. Using those keyword patterns, a clustering algorithm such as K-Means automatically grouped articles that used similar vocabulary. Once the clusters were created, I plotted the articles to help visualize how the groups were separated.

3. Explain the Graph

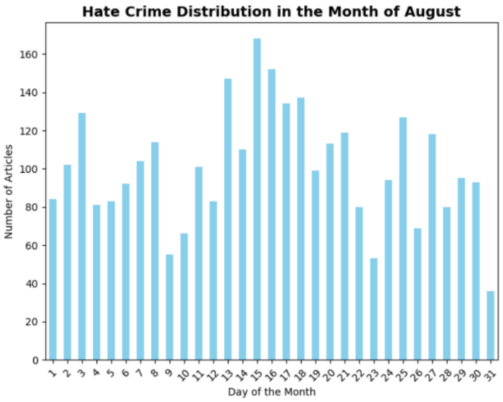
The graph for this question is a scatterplot that displays the clusters visually. Each point represents a single article, and the colors indicate which cluster the algorithm assigned it to. The x-axis and y-axis do not represent specific variables but rather two dimensions selected to make the clustering easier to view.

4. Harvest Highlights

The clustering analysis reveals that hate crime articles lack the distinct textual patterns needed for meaningful unsupervised grouping, as evidenced by the extremely low silhouette score of 0.014 indicating virtually no separation between proposed clusters. This suggests hate crime reporting employs overlapping vocabulary and narrative frameworks across different incident types, with articles sharing too much common language around law enforcement, community impact, and basic crime terminology to form natural thematic clusters.

[Part2:] Katelyn Wildermuth

[Question4]: What is the distribution of hate-crime reporting during the peak month?



1. **<u>Purpose</u>**

The purpose of this analysis is to determine when during the peak month of hate crime reporting was the majority of the reporting done. By concentrating on the number of reports occurring per day this allows stakeholders to determine when during the month is best to increase awareness of these patterns.

2. **<u>Methodology</u>**

This analysis used a cleaned data set that extracted a set date range and the count of article occurrences to accurately determine how many articles were published each day of the month. These values were used to determine where the peaks were during the month.

3. **<u>Explain the graph</u>**

The graph is a bar chart showing visually the distribution of hate crime reporting through the month of August. The x axis is labeled Day of the Month with each value being a day of the month. The y axis is labeled Number of Articles with each value being the number of articles published.

4. **<u>Harvest Highlights</u>**

The graph indicates that there are three distinct peaks of reporting during the month, indicating that hate crime reporting seems to follow a cyclical trend as the month progresses.

Overview

Our analysis reveals three key insights into hate crime reporting patterns. First, reporting peaks significantly in August, accounting for 20% of annual coverage and indicating seasonal trends. Second, coverage concentrates heavily in New York and California, suggesting regional disparities in incidents or media focus. Third, while decision trees successfully classify incident types from text, hate crime articles share too much overlapping vocabulary to form distinct clusters through unsupervised methods.

Contributions

Eva Santana:

My responsibility to this project was analyzing hate crime data through three key approaches. For Question 1, I used a **Decision Tree classifier** to predict incident types based on text features, using a text preprocessing technique to extract meaningful patterns from article summaries.

For Question 2, I conducted **temporal and geographic analysis** using time series visualization and state-level aggregation to examine reporting patterns across seasons and regions. The methodology involved date parsing, geographic mapping, and comparative analysis to identify hotspots and seasonal trends.

For Question 3, I implemented **K-Means clustering** with TF-IDF vectorization to discover natural groupings in the data. This involved dimensionality reduction using PCA for visualization and silhouette scoring to evaluate cluster quality. The significant challenge was the low silhouette score (0.014), indicating poor cluster separation, which revealed that hate crime articles share too much overlapping vocabulary to form distinct natural clusters.

Katelyn Wildemuth:

My responsibility for this project was analyzing a more concentrated trend within the hate crime data. For Question 4, I conducted a **temporal analysis** of the peak of reporting to analyze patterns across the month to determine when the most hate crimes were reported. This allowed for cyclical trends through the month to be shown.