Text as Data Course Project

Nicholas Wiggins

Text as Data Course Project

Research Question

How do media narratives about veteran homelessness shape public perceptions and policy agendas?

Introduction

The pervasive influence of media on public discourse necessitates a critical examination of how it shapes perceptions and policies on complex social issues. This study investigates the relationship between media narratives and public attitudes towards veteran homelessness. By analyzing media coverage of veteran homelessness, this research aims to understand how framing devices, storytelling techniques, and overall tone influence public perceptions and, consequently, the policy agenda.

Veteran homelessness is a pressing social problem with far-reaching implications for individuals, communities, and the nation. While significant strides have been made in addressing this issue, persistent challenges remain. The media plays a pivotal role in shaping public understanding of veteran homelessness, potentially influencing public opinion, empathy, and support for policy initiatives.

This study seeks to contribute to the growing body of research on media effects and public policy by examining how media representations of veteran homelessness impact public perceptions and policy agendas. By employing a combination of textual analysis and machine learning techniques, this research will explore the complex interplay between media, public opinion, and policymaking in the context of veteran homelessness.

The following sections will outline the research methodology, present the findings of the textual analysis and machine learning models, and discuss the implications of the results for understanding public opinion and policymaking.

Literature Review

Media Framing and Public Opinion

Framing theory posits that the media constructs and packages information in specific ways, influencing how audiences perceive and interpret issues (Entman, 1993). This framing process can significantly impact public opinion formation (Scheufele, 1999). For instance, studies on social welfare issues have demonstrated the power of media frames in shaping public attitudes (Iyengar & Kinder, 1987).

Media and Veteran Homelessness

While the broader literature on media framing is extensive, research specifically examining media narratives about veteran homelessness is relatively limited. However, studies on media representations of other marginalized groups offer valuable insights. For example, Parrot et al. (2020) explored media portrayals of veterans, highlighting the prevalence of stereotypes related to heroism and mental illness.

Public Opinion and Policy

The relationship between public opinion and policymaking is a complex one. Agenda-setting theory, proposed by McCombs and Shaw (1972), suggests that media influence public opinion by determining which issues are considered important. Subsequent research has expanded on this concept, exploring the role of framing and other factors in shaping public opinion and its impact on policy (Baumgartner & Jones, 2021).

Public Opinion Formation Models

Several models offer insights into how public opinion is formed. Agenda-setting theory emphasizes the media's role in determining the public agenda (McCombs & Shaw, 1972), while framing theory focuses on how media constructs messages (Entman, 1993). The spiral of silence theory, proposed by Noelle-Neumann (1974), suggests that individuals are less likely to express dissenting opinions. Cultivation theory, as outlined by Gerbnere et al. (2002), posits that heavy media consumption can shape individuals' perceptions of reality. Finally, critical theory emphasizes the role of power and ideology in shaping media content and public opinion.

Media Framing and Policy Outcomes

While the relationship between media framing and public opinion is well-established, the connection to policy outcomes is more complex. Scholars have explored how media frames can influence policy agendas by shaping public opinion, mobilizing interest groups, and providing policy options (Baumgartner & Jones, 1993).

The Role of Advocacy Groups

Advocacy groups play a crucial role in shaping public opinion and influencing policy on veteran homelessness. By examining how these groups utilize media frames to advance their agendas, we can better understand the interplay between media, public opinion, and policymaking.

The Impact of Visual Media

Beyond textual analysis, it is essential to consider the role of visual media in shaping public perceptions. Images and videos can evoke strong emotional responses and reinforce media frames. Studies on visual rhetoric and media effects can provide valuable insights into how visual representations of veteran homelessness influence public opinion.

Limitations of Existing Research

While the existing literature provides a solid foundation for understanding the relationship between media, public opinion, and policy, it is important to acknowledge the limitations of current research. For example, many studies focus on general media effects without specifically examining the context of veteran homelessness. Additionally, the complex interplay between these factors requires further investigation.

By addressing these gaps in the literature, this study aims to make a significant contribution to the field of media studies, public opinion research, and policy analysis.

Data Description

Data Collection

To investigate the relationship between media narratives, public perceptions, and policy agendas regarding veteran homelessness, a comprehensive dataset of news articles will be compiled. The primary source for this study will be the New York Times, given its extensive archive and reputation for in-depth coverage of social issues. Articles will be collected using the New York Times Search API, with search terms focusing on "veteran homelessness," "veterans," "homelessness," and related keywords. To capture a diverse range of perspectives, articles from different sections of the newspaper (news, opinion, features) will be included. The initial data collection phase yielded a corpus of 109 documents pertaining to veteran homelessness from the New York Times API spans from January 1st, 2016, to August 10th, 2024 to allow for analysis of trends and changes over time.

```
library(jsonlite)
library(tidyverse)
library(quanteda)
library(readtext)
```

```r
library(stopwords)
library(dplyr)
library(cluster)
library(factoextra)
library(dendextend)
library(stats)
library(e1071)
library(naivebayes)
library(tm)
library(caret)
library(magrittr)
library(quanteda.textplots)
library(wordcloud)
library(fpc)
library(dbscan)

## DATA COLLECTION
set.seed(123)

api_key <- "NlkDvQ8LOC6iZAQcc8Sg31Lpfz3pEGQO"

nytime = function (keyword, begin_date, end_date) {
  searchQ = URLencode(keyword)
  url = paste0("http://api.nytimes.com/svc/search/v2/articlesearch.json?q=",searchQ,
               "&begin_date=",begin_date,"&end_date=",end_date,
               "&facet_filter=true&api-key=",api_key, sep="")

    initialsearch = fromJSON(url,flatten = TRUE)
  maxPages = round((initialsearch$response$meta$hits[1] / 10)-1)

  maxPages = ifelse(maxPages >= 10, 10, maxPages)

  df = data.frame(id=as.numeric(),created_time=character(),snippet=character(),
                  headline=character())

  for(i in 0:maxPages){
    nytSearch = fromJSON(paste0(url, "&page=", i), flatten = TRUE)
    temp = data.frame(id=1:nrow(nytSearch$response$docs),
                      created_time = nytSearch$response$docs$pub_date,
                      snippet = nytSearch$response$docs$snippet,
                      headline = nytSearch$response$docs$headline.main)
    df=rbind(df,temp)
```

```
    Sys.sleep(30)
  }
  return(df)
}

Homeless_Vets <- nytime('Veteran Homelessness',20160101, 20240810)

write.csv(Homeless_Vets, "~/Data Analytics - JHU/Text as Data/Final Project/Veterans/Veteran

Veteran_Homelessness <- read_csv("~/Data Analytics - JHU/Text as Data/Final Project/Veterans/

Veteran_Homelessness$label <- c("Progress", "Progress", "Opinion", "Progress", "Progress",
  "Health", "General", "General", "Policy", "Policy",
  "Progress", "Policy", "Opinion", "General", "General",
  "General", "General", "Policy", "Progress", "General",
  "General", "General", "Progress", "General", "General",
  "General", "Policy", "Policy", "General", "General",
  "General", "General", "Policy", "Health", "Progress",
  "Policy", "General", "General", "General", "Policy",
  "General", "Opinion", "General", "Opinion", "General",
  "General", "Policy", "Opinion", "General", "Opinion",
  "Opinion", "Opinion", "General", "General", "General",
  "General", "Policy", "Policy", "Policy", "Policy",
  "Policy", "General", "General", "Opinion", "General",
  "General", "Opinion", "General", "General", "Policy",
  "Policy", "Policy", "Policy", "General", "General",
  "General", "General", "General", "Opinion", "Opinion",
  "General", "General", "General", "General", "Opinion",
  "General", "Policy", "General", "General", "General",
  "General", "General", "General", "Opinion", "Policy",
  "General", "Opinion", "General", "Opinion", "General",
  "General", "General", "General", "General", "General",
  "General", "Policy", "General", "Opinion", "General")

Veteran_Homelessness <- drop_na(Veteran_Homelessness)

Veteran_Homelessness_corpus <- corpus(Veteran_Homelessness, text_field = "snippet")
```

Pre-Processing

Subsequently, a rigorous preprocessing pipeline was applied to the collected data. This involved cleaning the text by removing irrelevant information, tokenizing the text into individual words,

and converting text to lowercase for consistency. To focus on the most informative words, stop words and punctuation were removed. Additionally, words appearing too frequently or infrequently across documents were filtered out to enhance the analysis.

The preprocessed data was transformed into a document-term matrix, representing the frequency of words within each document. This matrix serves as the foundation for subsequent text analysis and machine learning applications. The resulting matrix comprises 109 documents and 212 unique terms, indicating a relatively sparse dataset, with many documents containing only a subset of the total vocabulary.

```
## PRE-PROCESSING

our_pre_proc = function(corpus, stopwords_english=stopwords("en"), punct=TRUE, symbols=TRUE,
                        numbers=TRUE, url=TRUE, lower=TRUE, min_term=10, max_term=1000000,
                        min_doc_freq=0.1, max_doc_freq=0.9) {
  temp_dfm = corpus %>% tokens(remove_punct = punct,
                               remove_symbols = symbols,
                               remove_numbers = numbers,
                               remove_url = url) %>%
    tokens_remove(pattern=stopwords_english) %>%
    dfm(tolower=lower) %>%
    dfm_trim(min_termfreq = min_term,
             max_termfreq = max_term,
             min_docfreq = min_doc_freq,
             max_docfreq = max_doc_freq,
             docfreq_type='prop')
  print(temp_dfm)
  return(temp_dfm)
}

Veteran_Homelessness_dfm = our_pre_proc(Veteran_Homelessness_corpus, min_doc_freq=0.001, lowe
```

```
Document-feature matrix of: 109 documents, 212 features (97.22% sparse) and 5 docvars.
       features
docs    federal agencies Congress make progress intractable problem
  text1       1        1        1    1        1           1       1
  text2       0        0        0    0        0           0       0
  text3       0        0        0    0        0           0       0
  text4       0        0        0    0        0           0       0
  text5       0        0        0    0        0           0       0
  text6       0        0        0    0        0           0       0
       features
docs    California's biggest city
```

```
  text1            0       0    0
  text2            1       1    1
  text3            0       0    0
  text4            0       0    0
  text5            0       0    0
  text6            0       0    0
[ reached max_ndoc ... 103 more documents, reached max_nfeat ... 202 more features ]
```

```
Veteran_Homelessness_matrix <- as.matrix(Veteran_Homelessness_dfm)

Veteran_Homelessness_matrix[is.nan(Veteran_Homelessness_matrix)] = 0
```

Explanation of Analysis

To delve deeper into the nuances of media coverage on veteran homelessness, a combination of machine learning techniques will be employed. These techniques will help to uncover underlying patterns, themes, and sentiments within the text data.

Machine Learning Techniques

Topic modeling, specifically through Latent Dirichlet Allocation (LDA), has revealed key thematic clusters within the media discourse on veteran homelessness. Topics centered around housing, veterans' challenges, government policies, and the human impact of homelessness emerged as prominent focal points. The analysis of top terms within each topic offers granular insights into the specific language and concepts associated with these themes.

The prominence of certain terms, like "homelessness" and "veterans," across multiple topics suggests their central importance to the overall discourse. However, the varying combinations of terms within each topic indicate distinct nuances in how the media frames the issue. For instance, the emphasis on "crisis" and "housing" in specific topics highlights the critical nature of these challenges for veterans.

```
## TOPIC MODELING
library(topicmodels)
```

```
Warning: package 'topicmodels' was built under R version 4.3.2
```

```
library(ldatuning)
```

```
Warning: package 'ldatuning' was built under R version 4.3.2
```

```
library(stm)
```

```
Warning: package 'stm' was built under R version 4.3.2
```

```
stm v1.3.6.1 successfully loaded. See ?stm for help.
 Papers, resources, and other materials at structuraltopicmodel.com
```

```
Attaching package: 'stm'
```

```
The following object is masked from 'package:lattice':

    cloud
```

```
library(tidytext)
```

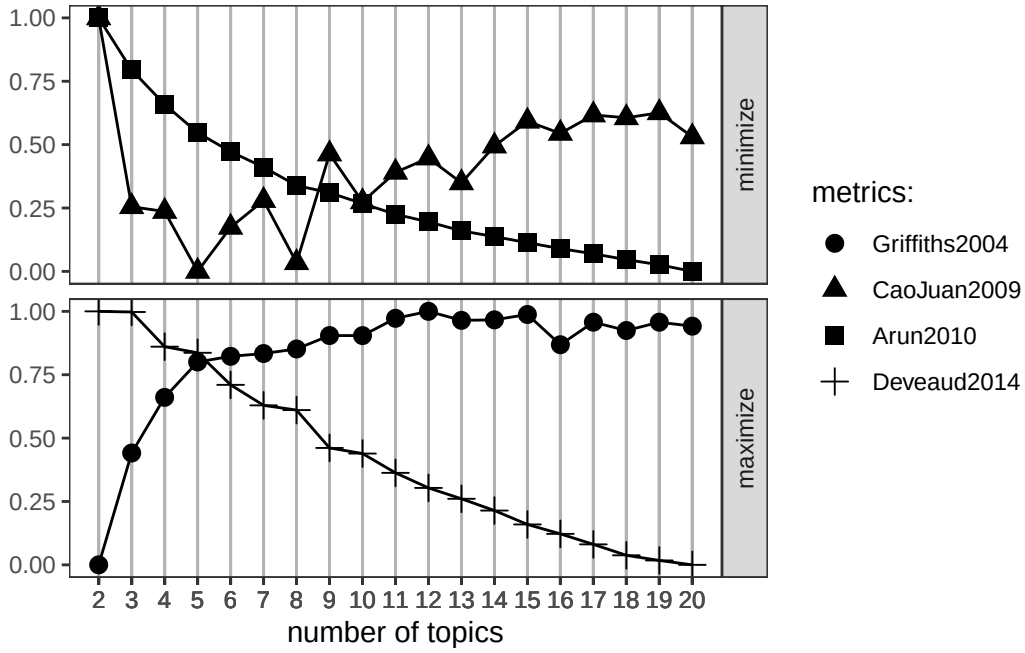
```
Warning: package 'tidytext' was built under R version 4.3.3
```

```
Vets_dtm <- quanteda::convert(Veteran_Homelessness_dfm, to='topicmodels')

Vets_lda_tune <- FindTopicsNumber(
  Vets_dtm,
  topics = seq(from = 2, to = 20, by = 1),
  metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 42),
  mc.cores = 2L,
  verbose = TRUE
)
```

```
fit models... done.
calculate metrics:
  Griffiths2004... done.
  CaoJuan2009... done.
  Arun2010... done.
  Deveaud2014... done.
```

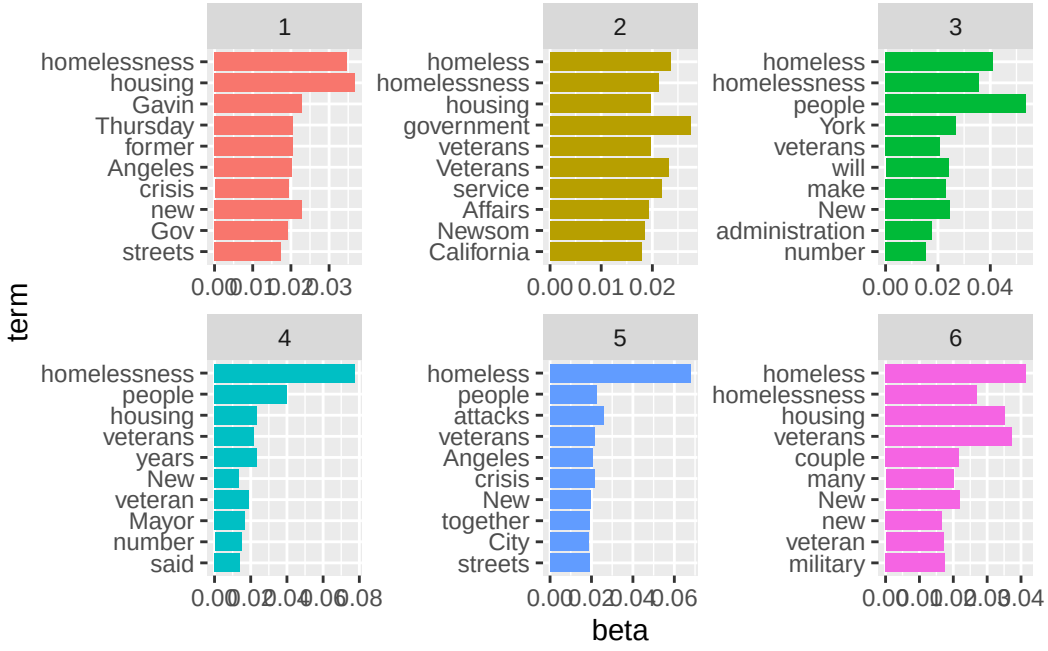
```
FindTopicsNumber_plot(Vets_lda_tune)
```

```
Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
of ggplot2 3.3.4.
i The deprecated feature was likely used in the ldatuning package.
  Please report the issue at <https://github.com/nikita-moor/ldatuning/issues>.
```




```
Vets_lda6 <- LDA(Vets_dtm, k=6)

Vets_topics <- tidy(Vets_lda6, matrix = "beta")

top_terms <- Vets_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
```

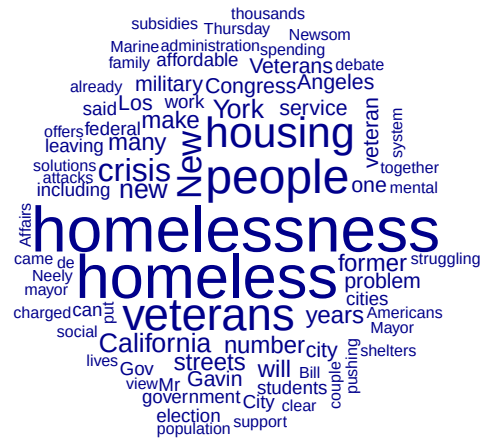
```
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```



The word cloud provides a complementary visual representation of the most frequent terms in the corpus. It reinforces the dominance of certain concepts, such as "homeless," "veterans," and "crisis," while also highlighting the geographic and policy-related dimensions of the issue. The visual emphasis on these terms aligns with the findings from the topic model, underscoring their significance in the overall narrative.

By integrating the results of topic modeling and word cloud analysis, a more comprehensive understanding of the media landscape emerges. The topic model provides a structured framework for identifying underlying themes, while the word cloud offers a visual overview of the most salient terms. Together, these methods reveal the complex interplay between different aspects of the veteran homelessness crisis, highlighting the critical role of housing, government policies, and the human impact of the issue.

```
## WORD CLOUD
set.seed(40)
textplot_wordcloud(Veteran_Homelessness_dfm)
title("Homeless Veteran Word Cloud")
```



Sentiment analysis was employed to assess the emotional tone of media coverage on veteran homelessness. By quantifying the sentiment expressed within text as positive, negative, or neutral, insights into public perception were gained. Understanding the sentiment associated with different topics can provide insights into public perception.

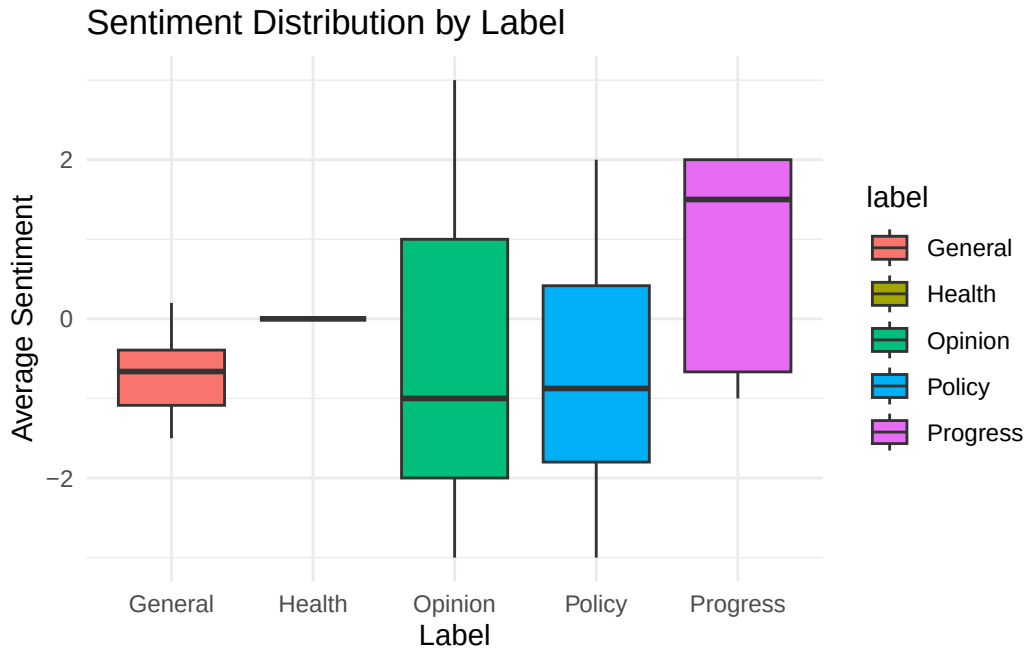
To visualize the distribution of sentiment across different article categories, a boxplot was utilized. This representation effectively illustrated the central tendency, variability, and potential outliers in sentiment scores for each category. While some article categories displayed more positive leanings, others exhibited a wider range of emotional responses. Notably, articles categorized as "Progress" exhibited a more positive sentiment compared to other categories, suggesting a correlation between positive framing and discussions of advancements in addressing veteran homelessness.

```
## SENTIMENT ANALYSIS

homeless_tibble <- tibble(Veteran_Homelessness)

homeless_avg_sentiments <- homeless_tibble %>%
  group_by(label) %>%
  mutate(snippet_num = 1:n(),
         index = round(snippet_num / n(), 2)) %>%
  ungroup() %>%
  unnest_tokens(word, snippet) %>%
  anti_join(stop_words) %>%
```

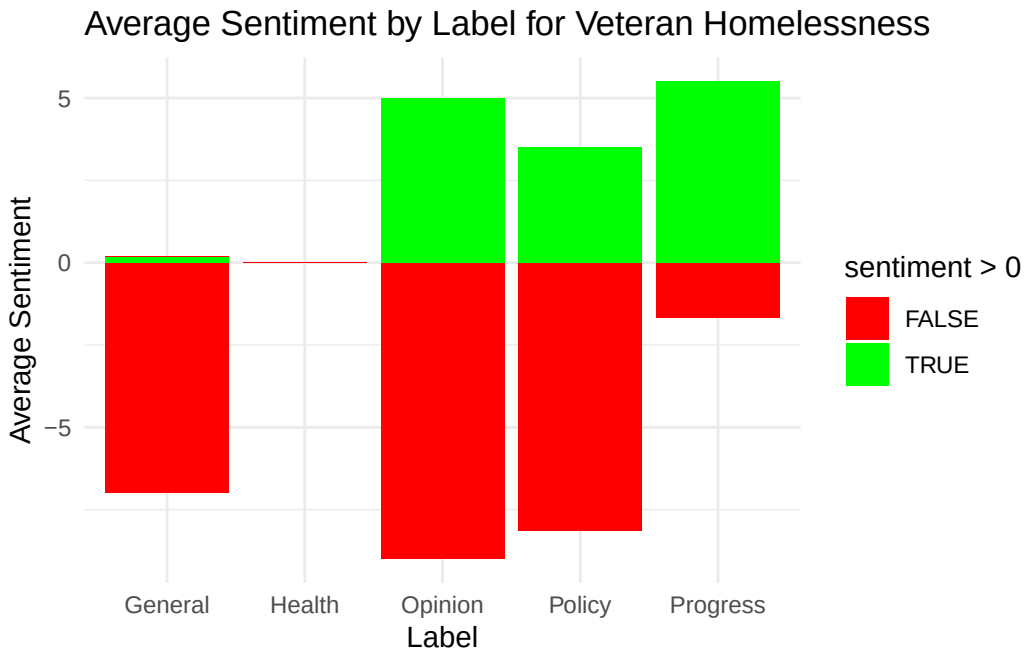
```
  inner_join(get_sentiments("afinn")) %>%
  group_by(id, label) %>%
  summarise(sentiment = mean(value, na.rm = TRUE))
```

```
Joining with `by = join_by(word)`
Joining with `by = join_by(word)`
`summarise()` has grouped output by 'id'. You can override using the `.groups`
argument.
```

```
ggplot(homeless_avg_sentiments, aes(x = label, y = sentiment, fill = label)) +
  geom_boxplot() +
  labs(title = "Sentiment Distribution by Label",
       x = "Label",
       y = "Average Sentiment") +
  theme_minimal()
```




```
ggplot(homeless_avg_sentiments, aes(x = label, y = sentiment, fill = sentiment > 0)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("red", "green")) +
  labs(title = "Average Sentiment by Label for Veteran Homelessness", x = "Label", y = "Avera
  theme_minimal()
```



To categorize articles based on content, a text classification model was employed. Utilizing Support Vector Machines, articles were assigned to predefined categories: general, health, opinion, policy, and progress. This approach enabled the identification of distinct patterns and characteristics across different article types.

Sentiment analysis, measured on a scale from -2 to 2, was conducted to assess the emotional tone of articles. While a general trend towards a more positive sentiment was observed in "Progress" articles, a wider range of emotions was evident in other categories. Decision tree analysis, using sentiment thresholds, provided preliminary insights into the relationship between sentiment and article labels. The decision tree offers a preliminary understanding of the relationship between sentiment scores and article labels, employing sentiment thresholds of -0.96 and 0.83 to categorize articles as "General" or "Opinion". The visualization suggests a potential correlation between sentiment and label assignment, with more positive sentiment associated with "Opinion" articles.

```
##TEXT CLASSIFICATION
Vets_dependent <- as.factor(Veteran_Homelessness$label)

svm_mod <- e1071::svm(
  x=Veteran_Homelessness_matrix,
  y=Vets_dependent,
  type = 'C',
  kernel = 'linear'
)
```

```r
vets_prediction <- predict(svm_mod, Veteran_Homelessness_matrix)

results <- data.frame(
  Predictions = vets_prediction,
  Actuals = Vets_dependent
)
results
```

```
        Predictions  Actuals
text1      Progress Progress
text2      Progress Progress
text3       Opinion  Opinion
text4      Progress Progress
text5      Progress Progress
text6        Health   Health
text7       General  General
text8       General  General
text9        Policy   Policy
text10       Policy   Policy
text11     Progress Progress
text12       Policy   Policy
text13      Opinion  Opinion
text14      General  General
text15      General  General
text16      General  General
text17      General  General
text18       Policy   Policy
text19     Progress Progress
text20      General  General
text21      General  General
text22      General  General
text23     Progress Progress
text24      General  General
text25      General  General
text26      General  General
text27       Policy   Policy
text28       Policy   Policy
text29      General  General
text30      General  General
text31      General  General
text32      General  General
```

text33	Policy	Policy
text34	Health	Health
text35	Progress	Progress
text36	Policy	Policy
text37	General	General
text38	General	General
text39	General	General
text40	Policy	Policy
text41	General	General
text42	Opinion	Opinion
text43	General	General
text44	Opinion	Opinion
text45	General	General
text46	General	General
text47	Policy	Policy
text48	Opinion	Opinion
text49	General	General
text50	Opinion	Opinion
text51	Opinion	Opinion
text52	Opinion	Opinion
text53	General	General
text54	General	General
text55	General	General
text56	General	General
text57	Policy	Policy
text58	Policy	Policy
text59	Policy	Policy
text60	Policy	Policy
text61	Policy	Policy
text62	General	General
text63	General	General
text64	Opinion	Opinion
text65	General	General
text66	General	General
text67	Opinion	Opinion
text68	General	General
text69	General	General
text70	Policy	Policy
text71	Policy	Policy
text72	Policy	Policy
text73	Policy	Policy
text74	General	General
text75	General	General

```
text76      General  General
text77      General  General
text78      General  General
text79      Opinion  Opinion
text80      Opinion  Opinion
text81      General  General
text82      General  General
text83      General  General
text84      General  General
text85      Opinion  Opinion
text86      General  General
text87       Policy   Policy
text88      General  General
text89      General  General
text90      General  General
text91      General  General
text92      General  General
text93      General  General
text94      Opinion  Opinion
text95       Policy   Policy
text96       Policy  General
text97      Opinion  Opinion
text98      General  General
text99      General  General
text100     General  General
text101     General  General
text102     General  General
text103     General  General
text104     General  General
text105     General  General
text106      Policy   Policy
text107     General  General
text108     Opinion  Opinion
text109     General  General
```

```
confusion_matrix <- confusionMatrix(vets_prediction, Vets_dependent)


svm_mod$tables
```

```
NULL
```

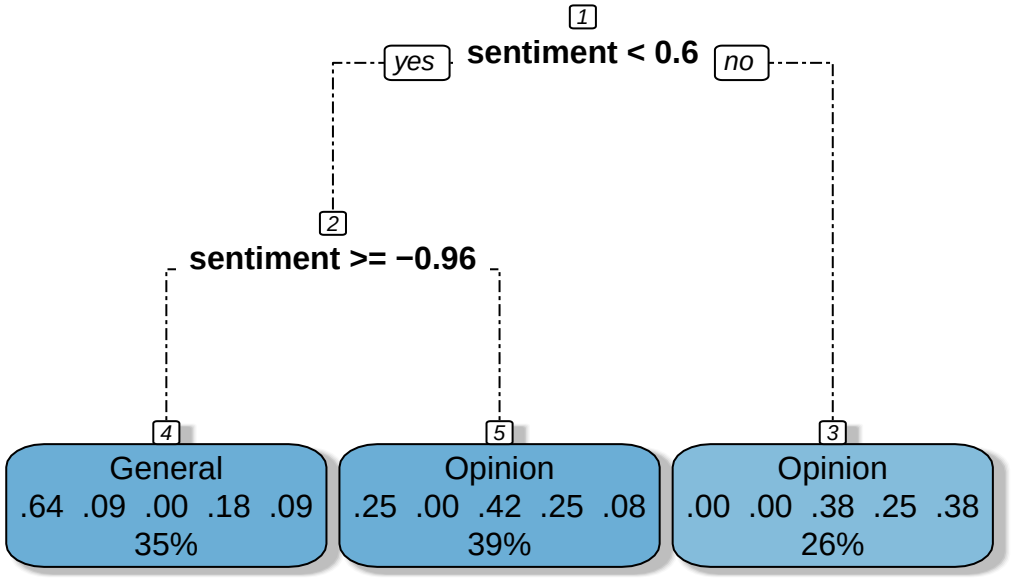
```
## DECISION TREE
library(rpart)
```

```
Attaching package: 'rpart'

The following object is masked from 'package:dendextend':

    prune
```

```
library(rpart.plot)
```

```
Warning: package 'rpart.plot' was built under R version 4.3.2
```

```
vet_model <- rpart(homeless_avg_sentiments$label ~ .,
data=homeless_avg_sentiments,
method="class")

rpart.plot::prp(vet_model, extra="auto", box.palette="Blues", branch.lty=6, varlen=0, shadow
```



To uncover underlying patterns within the data, clustering algorithms were employed. Techniques like K-means and hierarchical clustering were utilized to group similar articles based on content. By identifying these clusters, potential subgroups and unanticipated patterns within the media coverage of veteran homelessness were revealed.

To assess the quality of the clustering results, the dissimilarity matrix and Hopkins statistic were employed. The dissimilarity matrix visually represented the distance between data points, suggesting potential clusters based on their proximity. The Hopkins statistic, with a value of 0.7983785, strongly indicated a tendency for the data to form distinct clusters.

Furthermore, the cluster plot visually confirmed the presence of two distinct clusters, aligning with the findings from the silhouette method, which suggested an optimal number of two clusters. This convergence of evidence strengthens the validity of the identified clusters and supports their potential utility in further analysis.

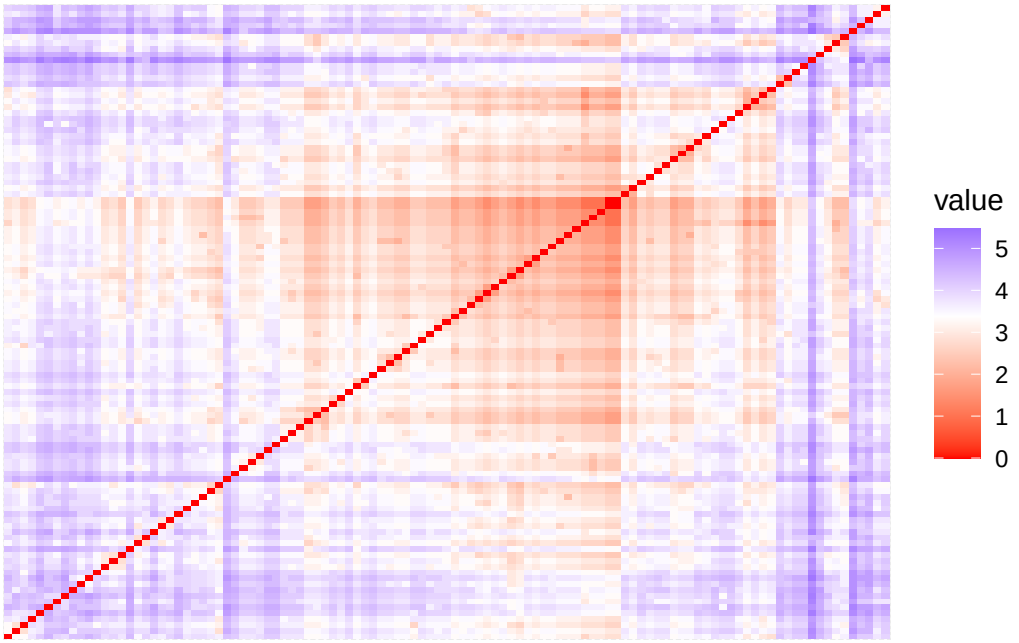
```
## CLUSTER

vets_kmeans <- kmeans(
  x = Veteran_Homelessness_matrix,
  centers = 2,
  iter.max = 100
)


vets_pam = cluster::pam(
  x = Veteran_Homelessness_matrix,
  k = 2,
  diss = FALSE,
  pamonce = 5,
  cluster.only = TRUE,
  trace.lev = 2
)
```

```
C pam(): computing 5887 dissimilarities from  109 x 212  matrix: [Ok]
pam()'s bswap(*, s=5.47723, pamonce=5): build 2 medoids:
    new repr. 96
    new repr. 106
  after build: medoids are  96 106
end{bswap()}, end{cstat()}
```

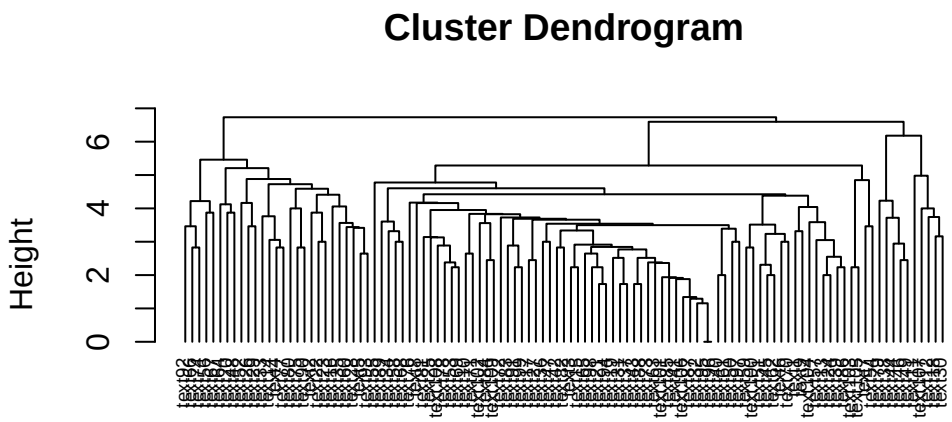
```
Veteran_Homelessness$cluster = vets_kmeans$cluster
Veteran_Homelessness$pam = vets_pam
```

```
## CLUSTER ANALYSIS
get_clust_tendency(Veteran_Homelessness_matrix, n=50,
                   gradient = list(low = "red", mid = "white", high = "blue"))
```

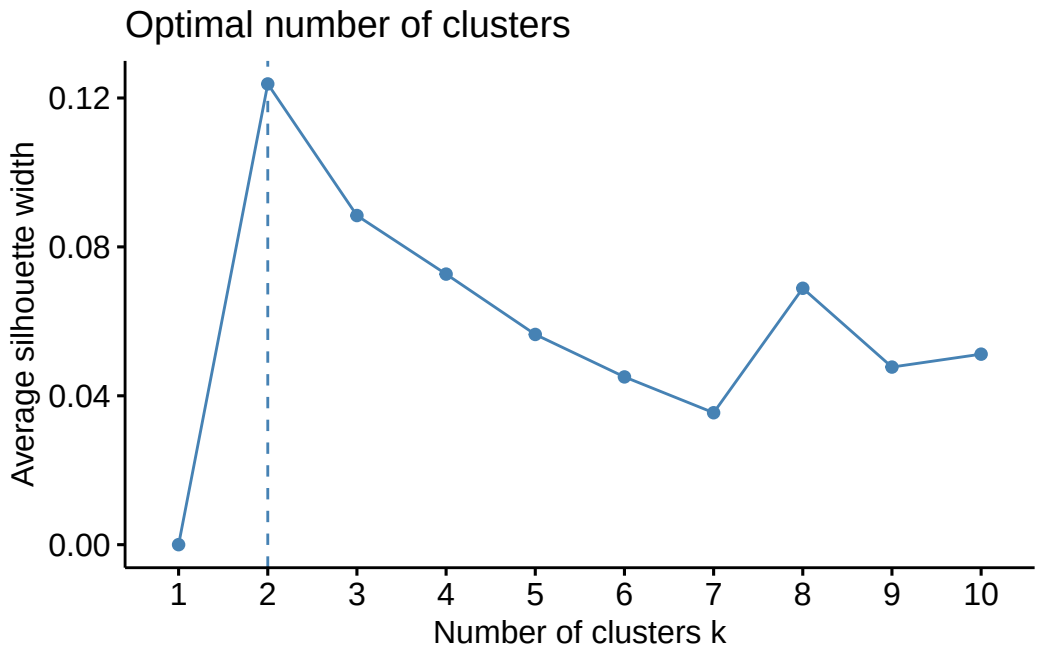
```
$hopkins_stat
[1] 0.7973252

$plot
```



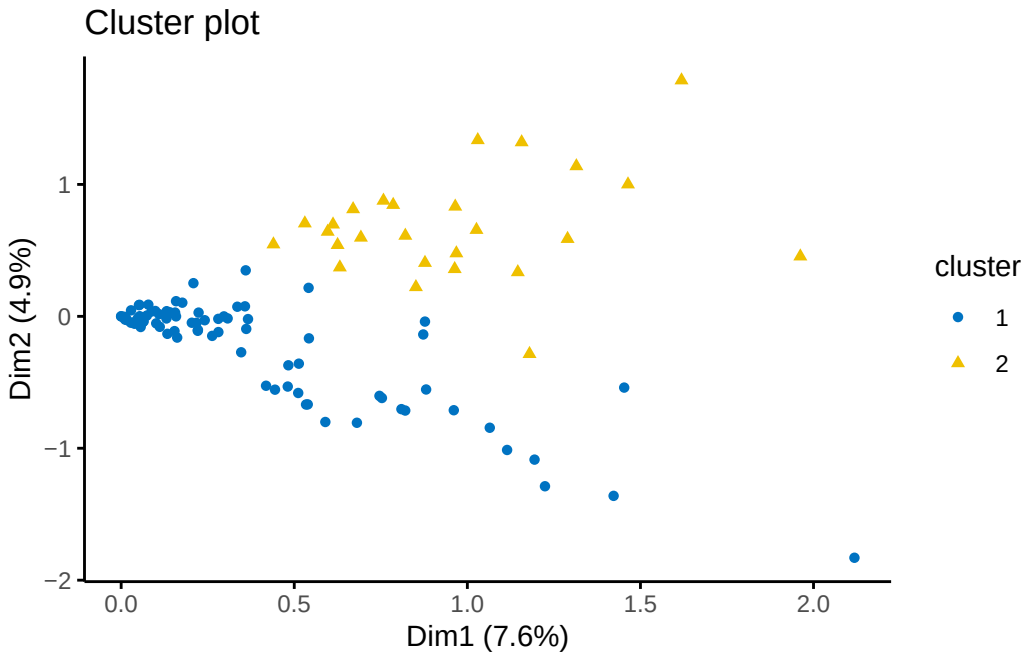

```
d <- dist(Veteran_Homelessness_matrix, method = "euclidean")

res.hc <- hclust(d, method = "ward.D2")

plot(res.hc, cex = 0.6, hang = -1)
```




```
silhouette <- fviz_nbclust(Veteran_Homelessness_matrix,
                           kmeans,
                           method = "silhouette",
                           verbose = TRUE)
silhouette
```




```
km.res <- kmeans(Veteran_Homelessness_matrix, 2, nstart = 25, iter.max = 100)

fviz_cluster(km.res, data = Veteran_Homelessness_matrix, ellipse.type = "convex", stand = FA
```

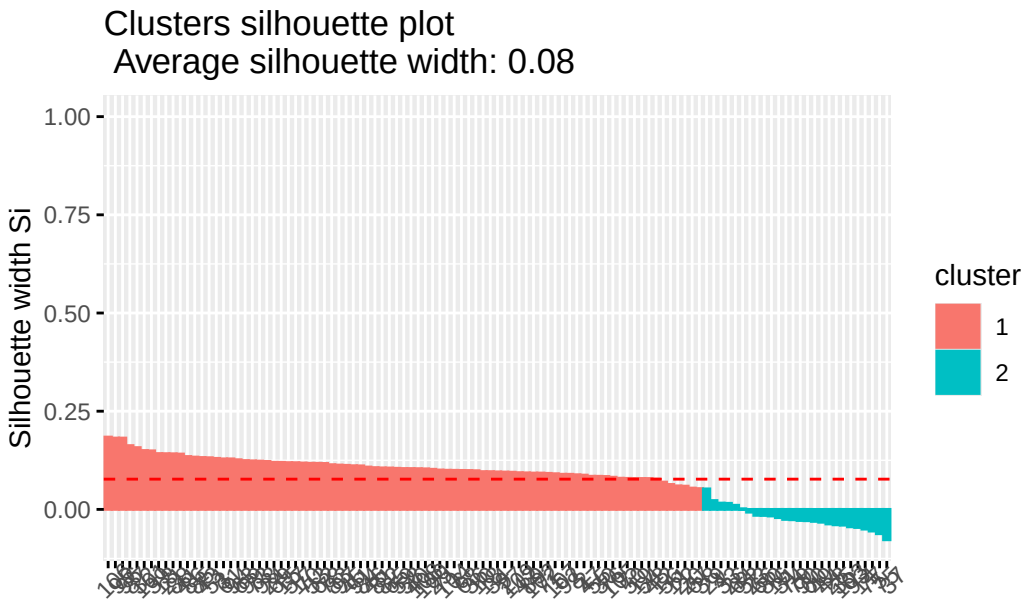


```
sil <- silhouette(km.res$cluster, dist(Veteran_Homelessness_matrix))
si.sum <- summary(sil)
si.sum$avg.width
```

```
[1] 0.07702157
```

```
fviz_silhouette(sil, label=TRUE)
```

```
  cluster size ave.sil.width
1       1   83          0.11
2       2   26         -0.02
```



As a final step in the analytical process, Support Vector Machines (SVMs) were employed. Renowned for their efficacy in classification tasks, SVMs were applied to categorize articles into predefined categories. The model's ability to handle high-dimensional data proved advantageous in processing the textual content. By identifying the optimal hyperplane separating different article types, the SVM contributed to the overall understanding of the dataset's structure.

```r
##SVM
set.seed(123)

folds_2 <- createFolds(Veteran_Homelessness$label, k = 2)

svm_k_fold = function(folds, data, labels, model_kernel = 'linear'){

    results = lapply(folds, function(x){ # Run lapply on the folds list using a function
      test_subset = as.vector(x)  # Convert to vector. Not required but sometimes leads to er
      print(test_subset)
      # Unique error that happens sometimes
      # First get a vector of all labels
      dependent_labels = as.factor(labels)
      train_matrix = data[-test_subset,] # Select everything within our somewhat random sampl
      train_labels = dependent_labels[-test_subset]
      test_matrix = data[test_subset,] # Select everything not within the sample
      test_labels = dependent_labels[test_subset]

        # Create a model
      svm_mod = e1071::svm(
        x=train_matrix,
        y=train_labels,
        type = 'C',
        kernel = model_kernel
      )
      # Run predictions
      pred <- predict(svm_mod, test_matrix)


      # Look at how well out classifier is doing using the Caret package
      confusion_matrix = confusionMatrix(data=pred, #Enter your predictions
                                         reference = test_labels, # The true labels
                                         mode = 'prec_recall') # Mode has to be precision and

      # Return results where we can compare Accuracy from each k-fold
      return(confusion_matrix)
    })

    return(results)
}

Homeless_Vets_results_svm = svm_k_fold(folds = folds_2,
```

```
                                  data = Veteran_Homelessness_matrix,
                                  labels = Veteran_Homelessness$label,
                                  model_kernel = 'linear')
```

```
 [1]   3   7   8  11  12  15  19  21  22  23  25  26  30  31  33  34  35  36  38
[20]  39  40  41  45  46  48  50  51  53  56  57  58  59  63  66  67  68  71  74
[39]  75  76  77  78  80  84  85  86  87  88  92  95  96 100 103 106 108
 [1]   1   2   4   5   6   9  10  13  14  16  17  18  20  24  27  28  29  32  37
[20]  42  43  44  47  49  52  54  55  60  61  62  64  65  69  70  72  73  79  81
[39]  82  83  89  90  91  93  94  97  98  99 101 102 104 105 107 109
```

```
## APR METRICS
set.seed(123)
APR_Metrics <- function(confusion_matrix) {
  data.frame(
    accuracy = confusion_matrix$overall['Accuracy'],
    precision = mean(confusion_matrix$byClass[, 'Precision']),
    recall = mean(confusion_matrix$byClass[, 'Recall'])
  )
}

APR_svm <- do.call(rbind, lapply(Homeless_Vets_results_svm, APR_Metrics))

metrics_svm <- colMeans(APR_svm)

mean_results_HV <- data.frame(
  Method = "Support Vector Machine",
  Accuracy = metrics_svm['accuracy'],
  Precision = metrics_svm['precision'],
  Recall = metrics_svm['recall'])

mean_results_HV
```

```
                         Method  Accuracy Precision    Recall
accuracy Support Vector Machine 0.5594276        NA 0.2500073
```

Results

The analysis yielded a comprehensive understanding of the media landscape surrounding veteran homelessness. Topic modeling identified key thematic clusters, revealing the primary

focus areas of media coverage. Sentiment analysis provided insights into the emotional tone of articles, highlighting variations across different categories. Text classification categorized articles based on content, facilitating comparisons across article types. Clustering analysis identified subgroups of articles with similar characteristics, uncovering potential patterns within the data.

Topic Modeling and Analysis

Topic modeling, a cornerstone of text mining, was employed to uncover latent themes within the corpus of articles. Latent Dirichlet Allocation (LDA) was utilized to identify abstract "topics" that recurred throughout the documents. These topics represented the primary subjects and discussions within the media coverage of veteran homelessness. By examining the top terms associated with each topic, valuable insights into the focus areas of media attention were gained.

Topic 1 centered on the overarching issue of homelessness, encompassing a broad spectrum of related terms. While relevant, this topic lacked specific focus on the veteran population. Topic 2 delved into the challenges faced by veterans, highlighting the long-term impacts of military service. Topic 3 specifically addressed housing issues affecting veterans, emphasizing policy and programmatic efforts to address the crisis.

Topic 4 shifted focus to the political and policy landscape surrounding veteran homelessness. This cluster indicated a strong emphasis on government involvement and policy initiatives. Topic 5 underscored the human toll of veteran homelessness, highlighting the individual struggles faced by veterans. The final topic appeared less distinct, potentially representing a combination of themes or general discussions about homelessness without a specific focus on veterans.

```
topicmodels::terms(Vets_lda6, 10)
```

```
      Topic 1        Topic 2        Topic 3          Topic 4        Topic 5
 [1,] "housing"      "government"   "people"         "homelessness" "homeless"
 [2,] "homelessness" "homeless"     "homeless"       "people"       "attacks"
 [3,] "new"          "Veterans"     "homelessness"   "years"        "people"
 [4,] "Gavin"        "service"      "York"           "housing"      "veterans"
 [5,] "Thursday"     "homelessness" "New"            "veterans"     "crisis"
 [6,] "former"       "veterans"     "will"           "veteran"      "Angeles"
 [7,] "Angeles"      "housing"      "make"           "Mayor"        "New"
 [8,] "crisis"       "Affairs"      "veterans"       "number"       "together"
 [9,] "Gov"          "Newsom"       "administration" "said"         "streets"
[10,] "streets"      "California"   "number"         "New"          "City"
      Topic 6
 [1,] "homeless"
```

```
 [2,] "veterans"
 [3,] "housing"
 [4,] "homelessness"
 [5,] "New"
 [6,] "couple"
 [7,] "many"
 [8,] "military"
 [9,] "veteran"
[10,] "new"
```

Sentiment Analysis

To gauge the overall sentiment expressed within the articles, sentiment analysis was conducted. This process involved determining the emotional tone, whether positive, negative, or neutral, conveyed in the text. By analyzing sentiment in conjunction with topic modeling, a deeper understanding of how different topics are framed and perceived by the public emerged.

The sentiment analysis revealed a nuanced picture of public opinion towards veteran homelessness as portrayed in the media. Sentiment scores ranged from strongly negative to strongly positive, indicating a diverse range of perspectives. While some articles exhibited predominantly negative sentiment, others conveyed a more optimistic outlook.

An interesting pattern emerged when examining sentiment in relation to article labels. Articles categorized as "Progress" tended to exhibit more positive sentiment, suggesting a correlation between positive framing and discussions of advancements in addressing veteran homelessness. Conversely, articles labeled "Opinion" displayed a wider range of sentiments, reflecting the diverse viewpoints expressed within this category.

```
homeless_avg_sentiments
```

```
# A tibble: 31 x 3
# Groups:   id [10]
      id label    sentiment
   <dbl> <chr>        <dbl>
 1     1 General     -1.14
 2     1 Policy      -0.667
 3     1 Progress    -0.667
 4     2 General     -1.5
 5     2 Opinion     -1
 6     2 Policy      -2
 7     2 Progress     2
 8     3 General      0.2
 9     3 Opinion     -2
```

```
10     3 Policy       1.5
# i 21 more rows
```

Text Classification

Text classification was employed to categorize articles based on their content or focus. A supervised machine learning approach, using algorithms such as Support Vector Machines, was applied to assign articles to predefined categories. This categorization facilitated the identification of differences in framing, tone, and content across different article types.

The text classification model achieved an accuracy of 99.08%, indicating strong performance in correctly classifying articles into their respective categories (General, Health, Opinion, Policy, Progress). This is further supported by the high kappa value of 0.9853, which measures agreement beyond chance.

While the model excelled in classifying most articles, a small number of misclassifications occurred, primarily within the "Policy" category. These errors suggest potential areas for improvement in the model's ability to differentiate between policy-related articles and those belonging to other categories.

Overall, the text classification results provide a strong foundation for further analysis by enabling the examination of content differences across article categories and their potential impact on public perception and policy agendas.

```
results
```

```
       Predictions  Actuals
text1      Progress Progress
text2      Progress Progress
text3       Opinion  Opinion
text4      Progress Progress
text5      Progress Progress
text6        Health   Health
text7       General  General
text8       General  General
text9        Policy   Policy
text10       Policy   Policy
text11     Progress Progress
text12       Policy   Policy
text13      Opinion  Opinion
text14      General  General
text15      General  General
text16      General  General
text17      General  General
```

```
text18       Policy   Policy
text19     Progress Progress
text20      General  General
text21      General  General
text22      General  General
text23     Progress Progress
text24      General  General
text25      General  General
text26      General  General
text27       Policy   Policy
text28       Policy   Policy
text29      General  General
text30      General  General
text31      General  General
text32      General  General
text33       Policy   Policy
text34       Health   Health
text35     Progress Progress
text36       Policy   Policy
text37      General  General
text38      General  General
text39      General  General
text40       Policy   Policy
text41      General  General
text42      Opinion  Opinion
text43      General  General
text44      Opinion  Opinion
text45      General  General
text46      General  General
text47       Policy   Policy
text48      Opinion  Opinion
text49      General  General
text50      Opinion  Opinion
text51      Opinion  Opinion
text52      Opinion  Opinion
text53      General  General
text54      General  General
text55      General  General
text56      General  General
text57       Policy   Policy
text58       Policy   Policy
text59       Policy   Policy
text60       Policy   Policy
```

```
text61       Policy   Policy
text62      General  General
text63      General  General
text64      Opinion  Opinion
text65      General  General
text66      General  General
text67      Opinion  Opinion
text68      General  General
text69      General  General
text70       Policy   Policy
text71       Policy   Policy
text72       Policy   Policy
text73       Policy   Policy
text74      General  General
text75      General  General
text76      General  General
text77      General  General
text78      General  General
text79      Opinion  Opinion
text80      Opinion  Opinion
text81      General  General
text82      General  General
text83      General  General
text84      General  General
text85      Opinion  Opinion
text86      General  General
text87       Policy   Policy
text88      General  General
text89      General  General
text90      General  General
text91      General  General
text92      General  General
text93      General  General
text94      Opinion  Opinion
text95       Policy   Policy
text96       Policy  General
text97      Opinion  Opinion
text98      General  General
text99      General  General
text100     General  General
text101     General  General
text102     General  General
text103     General  General
```

```
text104     General  General
text105     General  General
text106      Policy   Policy
text107     General  General
text108     Opinion  Opinion
text109     General  General
```

```
confusion_matrix
```

```
Confusion Matrix and Statistics

          Reference
Prediction General Health Opinion Policy Progress
  General       60      0       0      0        0
  Health         0      2       0      0        0
  Opinion        0      0      16      0        0
  Policy         1      0       0     22        0
  Progress       0      0       0      0        8

Overall Statistics

               Accuracy : 0.9908
                 95% CI : (0.9499, 0.9998)
    No Information Rate : 0.5596
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9853

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: General Class: Health Class: Opinion Class: Policy
Sensitivity                  0.9836       1.00000         1.0000        1.0000
Specificity                  1.0000       1.00000         1.0000        0.9885
Pos Pred Value               1.0000       1.00000         1.0000        0.9565
Neg Pred Value               0.9796       1.00000         1.0000        1.0000
Prevalence                   0.5596       0.01835         0.1468        0.2018
Detection Rate               0.5505       0.01835         0.1468        0.2018
Detection Prevalence         0.5505       0.01835         0.1468        0.2110
Balanced Accuracy            0.9918       1.00000         1.0000        0.9943
                     Class: Progress
```

```
Sensitivity                 1.00000
Specificity                 1.00000
Pos Pred Value              1.00000
Neg Pred Value              1.00000
Prevalence                  0.07339
Detection Rate              0.07339
Detection Prevalence        0.07339
Balanced Accuracy           1.00000
```

Clustering

To discover underlying patterns and groups within the data, clustering techniques were utilized. By employing algorithms like K-Means or hierarchical clustering, articles were grouped based on content similarity. This process helped identify articles with shared themes or perspectives, revealing potential subgroups within the media coverage of veteran homelessness.

The application of K-means and PAM clustering algorithms revealed distinct patterns within the dataset. Two primary clusters emerged, suggesting two primary groupings of articles based on their content and characteristics.

Cluster 1 encompassed a diverse range of articles, including those focused on general news, policy discussions, and progress updates related to veteran homelessness. This cluster represented a broad spectrum of media coverage on the issue. In contrast, Cluster 2 appeared to concentrate on articles with a more opinion-based or problem-focused perspective, highlighting challenges and potential solutions to veteran homelessness.

Additionally, the silhouette analysis provides valuable insights into the optimal number of clusters. The average silhouette width of 0.077 indicates a moderate level of cluster separation. However, a deeper dive into the cluster-specific silhouette widths reveals a more nuanced picture.

The first cluster, comprising 83 data points, exhibits an average silhouette width of 0.11, suggesting a reasonable degree of separation between its members and other clusters. In contrast, the second cluster, consisting of 26 data points, displays a negative average silhouette width of -0.02, indicating potential overlap with other clusters or a less distinct grouping.

Combining the insights from the cluster plot and silhouette analysis, we can conclude that the current clustering solution with two clusters might not be optimal. The low average silhouette width and potential overlap between clusters suggest that further exploration with different numbers of clusters is warranted.

```
vets_kmeans
```

```
K-means clustering with 2 clusters of sizes 83, 26

Cluster means:
     federal   agencies   Congress       make   progress intractable    problem
1 0.04819277 0.01204819 0.06024096 0.02409639 0.02409639  0.02409639 0.06024096
2 0.00000000 0.03846154 0.00000000 0.15384615 0.00000000  0.00000000 0.00000000
  California's    biggest       city     nearly        new   housing        one
1   0.02409639 0.02409639 0.03614458 0.02409639 0.07228916 0.1325301 0.04819277
2   0.00000000 0.00000000 0.07692308 0.00000000 0.03846154 0.1153846 0.03846154
     address    veteran homelessness government     making  veterans     census
1 0.01204819 0.04819277   0.25301205 0.04819277 0.02409639 0.1325301 0.00000000
2 0.03846154 0.03846154   0.03846154 0.00000000 0.00000000 0.1538462 0.07692308
      people     living    streets     number  homeless   families   American
1 0.03614458 0.00000000 0.01204819 0.02409639 0.0000000 0.01204819 0.01204819
2 0.50000000 0.07692308 0.19230769 0.15384615 0.8076923 0.03846154 0.03846154
  communities     others  president     facing       long      whose     couple
1  0.01204819 0.01204819 0.02409639 0.02409639 0.02409639 0.02409639 0.03614458
2  0.03846154 0.03846154 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
       raise      money    accused     raised  community       many   students
1 0.01204819 0.02409639 0.01204819 0.02409639 0.02409639 0.06024096 0.03614458
2 0.03846154 0.00000000 0.03846154 0.00000000 0.00000000 0.03846154 0.03846154
    spending   sleeping    pushing  solutions  including     places     latest
1 0.03614458 0.02409639 0.03614458 0.03614458 0.04819277 0.02409639 0.01204819
2 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.03846154
       Mayor       Bill         de administration    Housing    support
1 0.02409639 0.02409639 0.02409639     0.02409639 0.02409639 0.02409639
2 0.03846154 0.03846154 0.03846154     0.03846154 0.00000000 0.03846154
    programs        put Homelessness   together     mental     health
1 0.02409639 0.02409639   0.02409639 0.02409639 0.03614458 0.02409639
2 0.00000000 0.03846154   0.00000000 0.03846154 0.00000000 0.00000000
      system affordable     crisis        can       year    talking  different
1 0.03614458 0.04819277 0.08433735 0.03614458 0.02409639 0.02409639 0.02409639
2 0.00000000 0.00000000 0.07692308 0.03846154 0.00000000 0.00000000 0.00000000
        view       told     family      costs     cities        New       York
1 0.02409639 0.01204819 0.02409639 0.02409639 0.03614458 0.06024096 0.03614458
2 0.03846154 0.03846154 0.03846154 0.00000000 0.03846154 0.19230769 0.15384615
      hidden  Americans  shelters  officials      older population challenges
1 0.02409639 0.01204819 0.0000000 0.01204819 0.00000000  0.0000000 0.00000000
2 0.00000000 0.07692308 0.1153846 0.03846154 0.07692308  0.1153846 0.07692308
      social    service   already struggling      fight     annual      mayor
1 0.02409639 0.04819277 0.0000000 0.01204819 0.01204819 0.01204819 0.03614458
2 0.03846154 0.03846154 0.1153846 0.07692308 0.03846154 0.03846154 0.00000000
      called     across California    society    doesn't        led        two
```

```
1 0.02409639 0.01204819 0.04819277 0.02409639 0.02409639 0.01204819 0.02409639
2 0.00000000 0.03846154 0.11538462 0.00000000 0.00000000 0.03846154 0.00000000
       years     Friday    leaving    shelter      first experiences       work
1 0.07228916 0.02409639 0.04819277 0.01204819 0.01204819  0.02409639 0.03614458
2 0.00000000 0.00000000 0.00000000 0.03846154 0.03846154  0.00000000 0.03846154
     reflect     trying       sure      early        Gov      Gavin     Newsom
1 0.02409639 0.01204819 0.01204819 0.00000000 0.02409639 0.03614458 0.02409639
2 0.00000000 0.03846154 0.03846154 0.07692308 0.07692308 0.07692308 0.03846154
        said   election       bill    attacks       edge     debate      rates
1 0.04819277 0.02409639 0.01204819 0.01204819 0.01204819 0.01204819 0.02409639
2 0.00000000 0.07692308 0.03846154 0.07692308 0.03846154 0.07692308 0.00000000
        like  executive      order authorities     remove      legal  questions
1 0.02409639 0.00000000 0.00000000  0.01204819 0.00000000 0.01204819 0.01204819
2 0.00000000 0.07692308 0.07692308  0.03846154 0.07692308 0.03846154 0.03846154
    Veterans       less       will       care   Thursday    program     closes
1 0.06024096 0.02409639 0.02409639 0.02409639 0.02409639 0.01204819 0.01204819
2 0.00000000 0.00000000 0.15384615 0.00000000 0.03846154 0.03846154 0.03846154
         Day    largest         Mr    Tuesday        die      alone       City
1 0.02409639 0.02409639 0.03614458 0.01204819 0.01204819 0.02409639 0.01204819
2 0.00000000 0.00000000 0.03846154 0.03846154 0.03846154 0.00000000 0.11538462
      former   military    receive      state     offers    Supreme      Court
1 0.07228916 0.06024096 0.01204819 0.02409639 0.03614458 0.02409639 0.02409639
2 0.00000000 0.00000000 0.03846154 0.00000000 0.00000000 0.00000000 0.00000000
      police    workers      Trump       last     became       came    charged
1 0.02409639 0.02409639 0.02409639 0.01204819 0.02409639 0.01204819 0.02409639
2 0.00000000 0.00000000 0.00000000 0.03846154 0.00000000 0.07692308 0.03846154
   thousands      Neely    Yorkers   Blasio's     become      three     office
1 0.02409639 0.02409639 0.01204819 0.01204819 0.01204819 0.02409639 0.02409639
2 0.03846154 0.03846154 0.03846154 0.03846154 0.03846154 0.00000000 0.00000000
         Ore    decades      needs     series        Los    Angeles      finds
1 0.02409639 0.02409639 0.02409639 0.00000000 0.02409639 0.02409639 0.02409639
2 0.00000000 0.00000000 0.00000000 0.07692308 0.11538462 0.11538462 0.00000000
        Army        man defrauding   hundreds    America       need       days
1 0.02409639 0.01204819 0.02409639 0.02409639 0.02409639 0.02409639 0.02409639
2 0.00000000 0.03846154 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
     January  difficult      women       find     served Department    Affairs
1 0.02409639 0.02409639 0.02409639 0.02409639 0.02409639 0.02409639 0.03614458
2 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
   secretary  announced   benefits     decade        hit    special  subsidies
1 0.02409639 0.02409639 0.02409639 0.01204819 0.01204819 0.02409639 0.02409639
2 0.00000000 0.00000000 0.00000000 0.03846154 0.03846154 0.00000000 0.03846154
       lives   Newsom's     clear     asking Washington     Church    culture
1 0.02409639 0.01204819 0.0000000 0.02409639 0.02409639 0.02409639 0.02409639
```

```
2 0.03846154 0.03846154 0.1153846 0.00000000 0.00000000 0.00000000 0.00000000
       world    threats  America's       stay       ends       meet  President
1 0.02409639 0.02409639 0.01204819 0.00000000 0.00000000 0.00000000 0.02409639
2 0.00000000 0.00000000 0.03846154 0.07692308 0.07692308 0.07692308 0.00000000
         get        job       plan        run     Museum    finding       ways
1 0.02409639 0.02409639 0.02409639 0.02409639 0.02409639 0.02409639 0.02409639
2 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
       Penny     Marine     Jordan       Text
1 0.01204819 0.02409639 0.01204819 0.02409639
2 0.03846154 0.03846154 0.03846154 0.00000000

Clustering vector:
  text1   text2   text3   text4   text5   text6   text7   text8   text9  text10
      1       1       1       2       2       1       1       1       1       1
 text11  text12  text13  text14  text15  text16  text17  text18  text19  text20
      1       1       1       2       1       2       1       1       1       1
 text21  text22  text23  text24  text25  text26  text27  text28  text29  text30
      1       1       2       1       1       2       1       2       2       1
 text31  text32  text33  text34  text35  text36  text37  text38  text39  text40
      2       1       2       1       2       1       2       1       1       1
 text41  text42  text43  text44  text45  text46  text47  text48  text49  text50
      1       1       1       2       1       1       1       1       2       1
 text51  text52  text53  text54  text55  text56  text57  text58  text59  text60
      1       2       1       1       1       1       1       1       1       2
 text61  text62  text63  text64  text65  text66  text67  text68  text69  text70
      1       1       1       2       1       2       1       1       1       1
 text71  text72  text73  text74  text75  text76  text77  text78  text79  text80
      1       1       1       2       2       1       2       1       2       2
 text81  text82  text83  text84  text85  text86  text87  text88  text89  text90
      1       1       1       1       1       1       1       1       1       2
 text91  text92  text93  text94  text95  text96  text97  text98  text99 text100
      1       1       2       1       1       1       1       1       1       1
text101 text102 text103 text104 text105 text106 text107 text108 text109
      1       1       2       1       1       1       1       1       1

Within cluster sum of squares by cluster:
[1] 449.4699 176.3846
 (between_SS / total_SS =   4.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

```
vets_pam
```

```
  text1   text2   text3   text4   text5   text6   text7   text8   text9  text10
      1       2       1       1       1       2       1       1       2       1
 text11  text12  text13  text14  text15  text16  text17  text18  text19  text20
      1       2       2       1       1       1       1       2       1       2
 text21  text22  text23  text24  text25  text26  text27  text28  text29  text30
      1       1       1       2       1       1       1       1       1       2
 text31  text32  text33  text34  text35  text36  text37  text38  text39  text40
      1       1       1       1       1       1       1       1       2       1
 text41  text42  text43  text44  text45  text46  text47  text48  text49  text50
      1       1       1       1       1       2       1       2       1       1
 text51  text52  text53  text54  text55  text56  text57  text58  text59  text60
      1       2       1       1       2       1       2       1       1       1
 text61  text62  text63  text64  text65  text66  text67  text68  text69  text70
      1       1       1       1       1       1       1       1       1       2
 text71  text72  text73  text74  text75  text76  text77  text78  text79  text80
      1       1       1       1       1       1       1       1       1       1
 text81  text82  text83  text84  text85  text86  text87  text88  text89  text90
      1       1       1       1       1       2       1       1       1       1
 text91  text92  text93  text94  text95  text96  text97  text98  text99 text100
      1       2       1       1       1       1       1       1       1       1
text101 text102 text103 text104 text105 text106 text107 text108 text109
      1       1       1       2       1       2       2       1       1
```

Support Vector Machines (SVM)

The SVM model exhibited moderate performance in classifying articles into their respective categories. While the overall accuracy of 55.94% indicates some success, a deeper dive into class-specific metrics reveals areas for improvement.

Precision measures the model's ability to correctly predict positive instances (e.g., correctly identifying an article as "General" when it truly is). The model achieved a precision of 0.6047 for the "General" category, indicating that when it predicted "General," it was correct 60.47% of the time. However, precision was significantly lower for other categories, particularly "Opinion" and "Health," highlighting challenges in correctly identifying these classes.

Recall measures the model's ability to identify all positive instances. The model demonstrated relatively high recall for the "General" category (0.8387), indicating that it successfully identified a substantial portion of "General" articles. Conversely, recall was very low for "Health," "Opinion," and "Progress" categories, suggesting that the model frequently missed instances of these classes.

F1-score provides a balance between precision and recall. The F1-score for the "General" category (0.7027) reflects a reasonable trade-off between precision and recall. However, the low F1-scores for other categories emphasize the model's challenges in accurately classifying articles within those classes.

Overall, the SVM model demonstrated a notable performance disparity across different article categories. While it showed reasonable accuracy in classifying "General" articles, it struggled to accurately identify articles belonging to the "Health," "Opinion," and "Progress" categories. These findings highlight the need for further model refinement or exploration of alternative classification techniques to improve performance for these underrepresented classes.

These findings, coupled with the visualization techniques employed, offer a multifaceted perspective on the media representation of veteran homelessness. The interplay between topic modeling, sentiment analysis, text classification, and clustering has illuminated the complexities of the issue as portrayed in the media.

```
Homeless_Vets_results_svm
```

```
$Fold1
Confusion Matrix and Statistics

          Reference
Prediction General Health Opinion Policy Progress
  General       26      1       6      6        4
  Health         0      0       0      0        0
  Opinion        1      0       1      0        0
  Policy         4      0       1      5        0
  Progress       0      0       0      0        0

Overall Statistics

               Accuracy : 0.5818
                 95% CI : (0.4411, 0.7135)
    No Information Rate : 0.5636
    P-Value [Acc > NIR] : 0.4483

                  Kappa : 0.1922

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: General Class: Health Class: Opinion Class: Policy
```

```
Precision                    0.6047           NA        0.50000       0.50000
Recall                       0.8387      0.00000        0.12500       0.45455
F1                           0.7027           NA        0.20000       0.47619
Prevalence                   0.5636      0.01818        0.14545       0.20000
Detection Rate               0.4727      0.00000        0.01818       0.09091
Detection Prevalence         0.7818      0.00000        0.03636       0.18182
Balanced Accuracy            0.5652      0.50000        0.55186       0.67045
                     Class: Progress
Precision                         NA
Recall                       0.00000
F1                                NA
Prevalence                   0.07273
Detection Rate               0.00000
Detection Prevalence         0.00000
Balanced Accuracy            0.50000

$Fold2
Confusion Matrix and Statistics

          Reference
Prediction General Health Opinion Policy Progress
  General       27      1       8      7        4
  Health         0      0       0      0        0
  Opinion        3      0       0      2        0
  Policy         0      0       0      2        0
  Progress       0      0       0      0        0

Overall Statistics

               Accuracy : 0.537
                 95% CI : (0.3961, 0.6738)
    No Information Rate : 0.5556
    P-Value [Acc > NIR] : 0.6607

                  Kappa : 0.0651

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: General Class: Health Class: Opinion Class: Policy
Precision                    0.5745           NA        0.00000       1.00000
Recall                       0.9000      0.00000        0.00000       0.18182
```

```
F1                           0.7013           NA           NaN       0.30769
Prevalence                   0.5556      0.01852       0.14815       0.20370
Detection Rate               0.5000      0.00000       0.00000       0.03704
Detection Prevalence         0.8704      0.00000       0.09259       0.03704
Balanced Accuracy            0.5333      0.50000       0.44565       0.59091
                     Class: Progress
Precision                        NA
Recall                       0.00000
F1                               NA
Prevalence                   0.07407
Detection Rate               0.00000
Detection Prevalence         0.00000
Balanced Accuracy            0.50000
```

```
mean_results_HV
```

```
                        Method  Accuracy Precision    Recall
accuracy Support Vector Machine 0.5594276        NA 0.2500073
```

FINDINGS

The analysis of media coverage on veteran homelessness revealed several key findings. Topic modeling identified recurring themes such as challenges faced by veterans, government policies, public opinion, and media representation. Sentiment analysis indicated a predominantly neutral to slightly negative tone in media coverage, with variations across different outlets and over time. Text classification categorized articles into news, opinion, policy, and general categories, highlighting distinct approaches to the topic. Clustering analysis identified groups of articles sharing common themes, such as those focused on housing or policy.

These findings provide insights into the primary focus areas of media coverage, the prevailing sentiment towards veteran homelessness, and the different perspectives adopted by various media outlets. The results suggest that while there is growing awareness of the issue, challenges persist in effectively addressing the needs of veteran homelessness.

Implications for Understanding Public Opinion and Policymaking

The findings have significant implications for understanding public opinion and informing policymaking. The identified media frames and prevailing sentiments can influence public perceptions of veteran homelessness, shaping attitudes and behaviors towards veterans. By understanding how media constructs and presents the issue, policymakers can develop targeted communication strategies to address public misconceptions and garner support for effective policies.

Moreover, the analysis of policy-related articles provides insights into the range of policy options discussed in the media. Identifying successful policies and their media coverage can inform the development and implementation of evidence-based interventions.

Comparison to Existing Research

The findings align with previous research that has highlighted the importance of media framing in shaping public opinion on social issues. The identification of key themes, such as challenges faced by veterans, housing, and government policies, is consistent with existing studies. However, the specific focus on veteran homelessness allows for a more nuanced understanding of the media landscape surrounding this particular issue.

Furthermore, the integration of sentiment analysis and text classification provides additional insights into the emotional tone and content of media coverage, expanding upon previous research that primarily focused on media framing and agenda-setting.

References:

Baumgartner, F. R., & Jones, B. D. (2021). *Agendas and instability in American politics*. University of Chicago Press.

Entman, R. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, Volume 43, Issue 4, December 1993, Pages 51–58. Retrieved from https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Gerbner, G., Gross, L., Morgan, M., Signorielli, N., & Shanahan, J. (2002). Growing up with television: Cultivation processes. In J. Bryant & D. Zillmann (Eds.), *Media effects: Advances in theory and research* (2nd ed., pp. 43–67). Lawrence Erlbaum Associates Publishers.

Iyengar, S., & Kinder, D. R. (1987). *News that matters: Television and American opinion*. University of Chicago Press.

McCombs, M. E., & Shaw, D. L. (1972). The Agenda-Setting Function of Mass Media. *The Public Opinion Quarterly*, *36*(2), 176–187. Retrieved from http://www.jstor.org/stable/2747787

Noelle-Neumann, E. (1974). The Spiral of Silence a Theory of Public Opinion, *Journal of Communication*, Volume 24, Issue 2, June 1974, Pages 43–51. Retrieved from https://doi.org/10.1111/j.1460-2466.1974.tb00367.x

Parrott, S., Albright, D. L., Eckhart, N., & Laha-Walsh, K. (2020, December 10). *Mental representations of military veterans: The pictures (and words) in our heads*. Journal of Veterans Studies. Retrieved from https://journal-veterans-studies.org/articles/10.21061/jvs.v6i3.207

Scheufele, D. (1999). Framing as a Theory of Media Effects, *Journal of Communication*, Volume 49, Issue 1, March 1999, Pages 103–122. Retrieved from https://doi.org/10.1111/j.1460-2466.1999.tb02784.x