Capstone Project

Assignment 1

Course code: CSA1635

Course : Data Warehouse And Data Mining

S.No: 4

Name : Cherukuri sai murali

Reg. No: 192211951

Slot : A

Title : Sentiment Analysis for Political Opinion Mining using k-Nearest Neighbors (k-NN)

Assignment Release Date :

Assignment Preliminary Stage (Assignment 1) submission Date :

Mentor Name : MR. Raghavendran. R

Mentor Phone number and Department : 9087382665 ,PURE & APPLIED MATHEMATICS

1.Preliminary Stage

1.1 Assignment Description :

Sentiment Analysis for Political Opinion Mining using k-Nearest Neighbors (k-NN) entails preprocessing political texts to eliminate noise and standardize the data, extracting features using methods like word embeddings or Bag-of-Words, training the k-NN model on labeled data to classify sentiments (positive, negative, and neutral) based on the sentiments of its nearest neighbors, assessing model performance with metrics like accuracy and F1-score, adjusting hyperparameters like the number of neighbors (k), and utilizing the trained model for practical uses like social media sentiment monitoring and political discourse analysis. Stakeholders can learn more about the dynamics of public mood and opinion patterns related to political issues with this technique.

1.2 Assignment Work Distribution :

- **Project Scope Definition:**

The scope of the statement "Sentiment Analysis for Political Opinion Mining" project is to create a computer framework for analyzing and comprehending public opinion on political issues. In order to ascertain opinions held by people or groups about political issues, candidates, policies, or events, the project intends to gather and examine a wide variety of political texts, such as news articles, social media posts, and public statements. Through the analysis of public opinion toward political issues and people, the project seeks to advance informed governance and democratic engagement by offering insightful information for decision-making and comprehending societal trends. Through sentiment analysis of various textual data sources, its goals include finding patterns in sentiment, evaluating political environments, and influencing policy actions.

Sentiment analysis for political opinion mining using k-Nearest Neighbors (k-NN) has several specific goals. These include identifying influential features in political sentiment, optimizing k-NN parameters to improve sentiment classification accuracy, and deploying an effective and scalable model for real-

time monitoring of public sentiment on political issues. In the end, this model helps with informed policy formulation and decision-making.

Data Collection and Preparation:

1.Identify the Data Sources:

The data sources for Sentiment Analysis for Political Opinion Mining using k-Nearest Neighbors (k-NN) include social media platforms, news websites, and political blogs, offering diverse perspectives on public sentiment towards political topics.

2.Develop a Data Collection:

The data collection plan for Sentiment Analysis for Political Opinion Mining using k-Nearest Neighbors (k-NN) involves sourcing data from social media, news websites, and political blogs via web scraping and API access, supplemented by manual collection methods for transcripts and reports

3.Data Preprocessing:

the collected data goes through several data quality control steps. First, data cleansing is performed to eliminate any unwanted elements like unique characters or web addresses. Next, data preprocessing is done. It involves breaking the text into tokens, converting all words to lowercase, removing common words (stopwords), and stemming words to extract their root forms. This process standardizes the text and prepares it for feature extraction, making it ready for analysis.

4.Consistency:

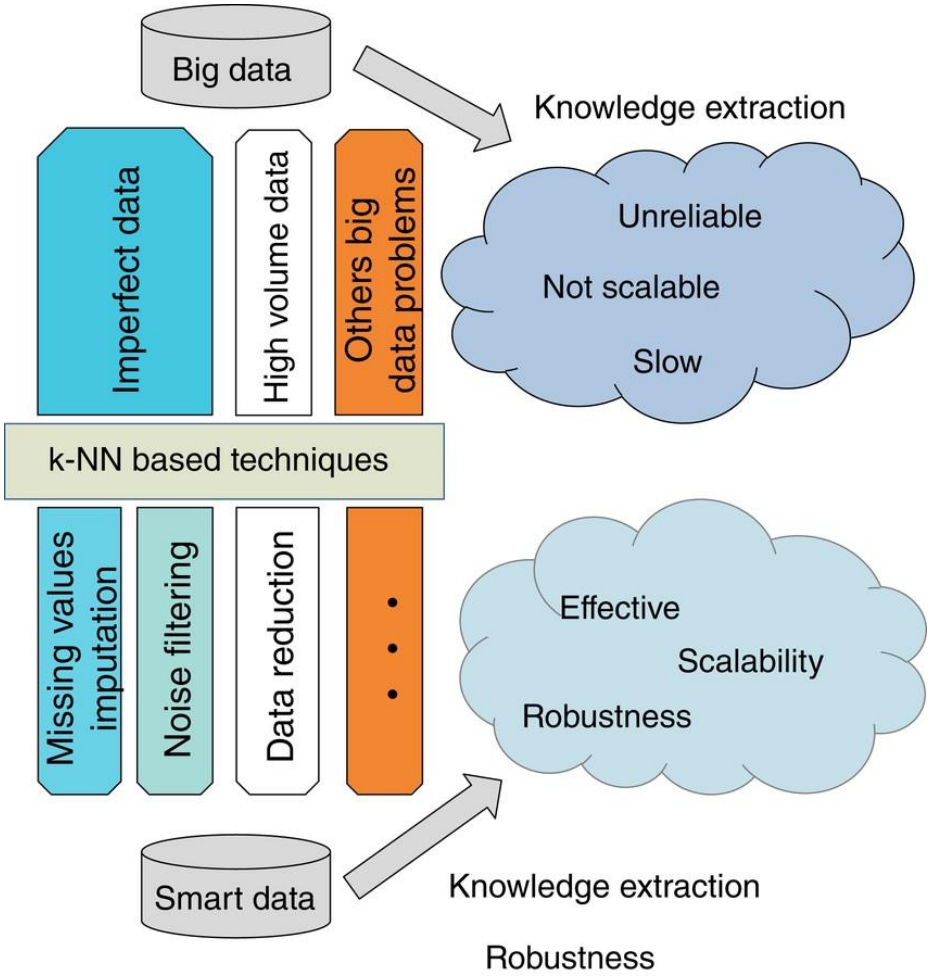
Transforming the data in this manner ensures consistency in its representation, leading to more accurate and effective sentiment analysis models.

EXPLORATORY DATA ANALYSIS:

data analysis for Sentiment Analysis for Political Opinion Mining using k-Nearest Neighbors (k-NN), assess the class distribution of sentiment labels and analyze word frequency and n-gram patterns to identify prevalent sentiments and key topics associated with political opinions. Visualization techniques like word clouds and temporal sentiment trends aid in uncovering insights crucial for model training and understanding public sentiment dynamics.

Sentiment dynamics on social media, topic-specific sentiment variations, and temporal shifts in sentiment are all examined in Sentiment Analysis for Political Opinion Mining using k-Nearest Neighbors (k-NN) in order to gain insights into public opinion dynamics and guide strategic decision-making in politics and governance. Through the identification of sentiment patterns across various settings and platforms, stakeholders can obtain significant insights into public sentiment trends and adjust their strategies accordingly for efficient communication and policy development.

VISUALIZE THE DATA USING ARCHITECTURE USING KNN-ALGORITHM.



PROBLEM STATEMENT:

The problem statement for Sentiment Analysis for Political Opinion Mining using k-Nearest Neighbors (k-NN) revolves around building a robust machine learning model capable of accurately classifying sentiments expressed in political texts. Key objectives include addressing noise inherent in textual data, tackling context-dependent sentiment interpretation, and enabling real-time analysis to provide timely insights into public opinion dynamics. This entails preprocessing data to remove noise and standardize text representations, extracting meaningful features from political texts, and training a k-NN model to classify sentiments based on neighboring data points. Challenges such as handling sarcasm, irony, and nuanced language expressions in political discourse must be overcome to ensure the model's effectiveness. Additionally, ensuring scalability and efficiency for processing large volumes of data from diverse sources is crucial for practical deployment in real-world applications. Ultimately, the goal is to develop a reliable sentiment analysis tool that aids stakeholders, including policymakers, political analysts, and campaign strategists, in understanding and responding to public sentiment trends on political issues.

ABSTRACT:

Sentiment analysis for political opinion mining using k-Nearest Neighbors (k-NN) offers a way for examining the opinions shared in political texts that are gathered from many sources, including news articles, social media, and public remarks. The main objective of the project is to create a machine learning model that employs the k-NN method to categorize attitudes into three groups: positive, negative, and neutral. The study tackles issues including sentiment interpretation that is context-dependent, noise in textual data, and the requirement for real-time analysis to yield meaningful insights into the dynamics of public opinion on political matters. The project intends to provide an efficient sentiment analysis tool through feature extraction techniques to represent textual information numerically, preprocessing strategies to clean and standardize data, and k-NN model training.

KEY WORDS: Sentiment Analysis, Political Opinion, Mining, k-Nearest Neighbors

PROPOSED DESIGN WORKS:

1.Data Collection:

Data Collection for Sentiment Analysis for Political Opinion Mining using k-Nearest Neighbors (k-NN) involves gathering textual data from various sources to analyze sentiments towards political topics. The process begins by identifying relevant platforms such as social media (e.g., Twitter, Facebook), news websites, political blogs, and public statements. Web scraping techniques and APIs are utilized to collect data, ensuring a diverse range of political opinions and viewpoints.

2.Data processing:

Sentiment analysis for Political Opinion Mining utilizing k-Nearest Neighbors (k-NN) data processing requires a number of procedures to get the gathered textual data ready for analysis. First, to clean up the data, noise reduction techniques such removing punctuation, special characters, and unnecessary information are used.

3. K-Nearest Neighbours Algorithm:

The k-Nearest Neighbors (k-NN) algorithm is a simple and intuitive machine learning approach used for classification and regression tasks. In the context of sentiment analysis for political opinion mining, k-NN is employed to classify the sentiment of political texts based on the sentiments of neighboring data points.

4. Association Rule:

Association rule mining is not typically directly applied to sentiment analysis tasks like political opinion mining; however, it can be used to uncover patterns between sentiment, political topics, or other factors when combined with sentiment analysis results. By identifying associations between sentiment and

political attributes, association rule mining offers insights that can inform decision-making processes and further analysis in political contexts.

FUNCTIONALITY:

The functionality of Sentiment Analysis for Political Opinion Mining using k-Nearest Neighbors (k-NN) involves preprocessing political text data, training a k-NN model to classify sentiments, and providing insights into public opinion on political issues. Key features include data collection from various sources, data preprocessing to remove noise and standardize text, feature extraction using techniques like Bag-of-Words or TF-IDF, model training to classify sentiments based on neighboring data points, and evaluation of model performance. The system facilitates informed decision-making for policymakers, political analysts, and campaign strategists by offering timely insights into public sentiment trends and sentiment dynamics surrounding political issues and figures.

ARCHITECTURE DESIGN:

Sentiment analysis for Political Opinion Mining using k-Nearest Neighbors (k-NN) is an architecture that includes gathering data from various sources, standardizing text through preprocessing, extracting features for numerical representation, training the model using the k-NN algorithm, and assessing the model's performance. Deployment makes real-world uses possible, such tracking public opinion and assisting political campaigns, after certification. Monitoring and retraining on a regular basis guarantee sustained performance. This modular method makes it easier to analyze political text data effectively and offers insightful information about how the public feels about various political problems.

LAYOUT AND DESIGN:

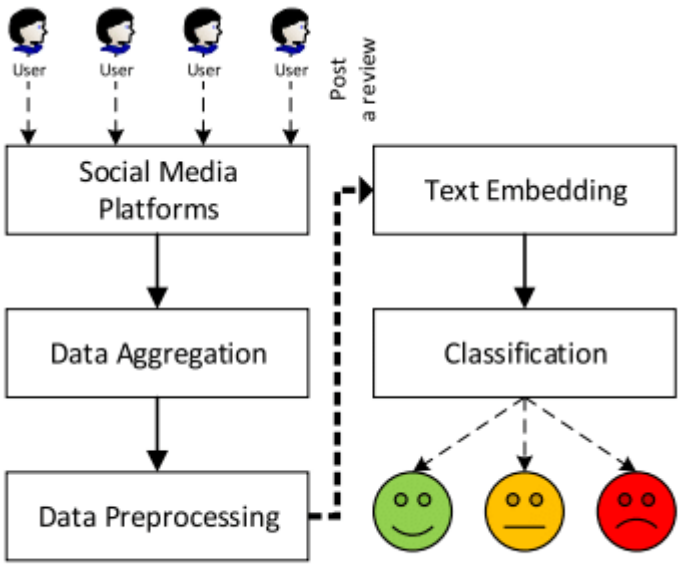


FIGURE 1. Basic steps of sentiment analysis on social media.

1.FLEXIBLE LAYOUT:

Flexible layout refers to the ability to adapt the layout or structure of the sentiment analysis system to different requirements or constraints. This could involve adjusting the feature representation of the text data, experimenting with different distance metrics, or fine-tuning the parameters of the k-NN algorithm to achieve the best performance for a particular task or dataset.

2.USER FRIENDLY:

Political opinion mining is a crucial aspect of understanding public sentiment towards various political issues, candidates, or parties. Employing techniques such as sentiment analysis using k-Nearest Neighbors (k-NN) can provide valuable insights into the prevailing attitudes and emotions within society. To ensure user-friendliness in such an endeavor, it's essential to create a streamlined interface where users can effortlessly input text and receive sentiment analysis results.

3.COLOUR SELECTION:

For color selection, k-NN can also be used, but it's not a typical application. In this context, you could potentially use k-NN to classify colors based on their RGB values or other color representations. However, other methods, such as clustering algorithms, are more commonly used for color selection tasks.

FUNCTIONS:

Sentiment analysis for political opinion mining involves analyzing text data to determine the sentiment expressed towards political figures, policies, or events. One approach to performing sentiment analysis is using the k-Nearest Neighbors (k-NN) algorithm. In this method, a dataset of political texts or social media posts, along with their corresponding sentiment labels, is collected and preprocessed by removing noise and transforming the text into numerical feature vectors, often using techniques like TF-IDF. The k-NN model is then trained on this preprocessed data, learning to classify new instances based on the similarity of their features to those of the labeled instances in the training set. After training, the model is evaluated using metrics such as accuracy, precision, recall, and F1-score on a separate validation or test dataset. Fine-tuning of hyperparameters, such as the number of neighbors (k), may be performed to optimize performance. Once trained and evaluated, the k-NN model can be used to predict the sentiment of new political texts or social media posts by transforming them into feature vectors and classifying them based on their similarity to the training data.

FEASIBLE ELEMENTS USED:

Sentiment analysis for political opinion mining utilizing k-Nearest Neighbors (k-NN) employs feasible elements that include text preparation methods such as tokenization and stopword removal to clean up data. Text can be converted into numerical representations that are appropriate for model training using feature extraction techniques like TF-IDF and Bag-of-Words. For sentiment classification, the k-NN algorithm which is well-known for being straightforward and efficient with datasets of a reasonable size is utilized. Real-world information from many sources, such as news websites and social media, offers pertinent political text samples for analysis. Last but not least, scalable deployment techniques and monitoring procedures guarantee the sentiment analysis system's usefulness and continuous efficacy.

LOGIN TEMPLATE:

1.login process:

User enter their username and password to access the system.

2.Sign up process:

New user can create an account by providing their details and creating a password.

3.Other templates:

Include profile management, data settings and system preferences.

CONCLUSION:

Sentiment analysis for political opinion mining using k-Nearest Neighbors (k-NN) provides a useful method for figuring out how the public feels about political individuals and issues. This methodology allows reliable sentiment categorization of political writings obtained from various platforms by utilizing practical components including the k-NN algorithm, feature extraction techniques, and text preprocessing techniques. Efficient data processing, model training, and deployment for practical applications are made possible by the modular architecture design. Stakeholders can learn about the fluctuations of public opinion and make wise judgments in political situations by implementing thorough evaluation and monitoring procedures. New developments in sentiment analysis techniques could improve our understanding of public sentiment and its implications for democracy and government as technologydevelopsfurther.

ASSIGNMENT 2

R-PROGRAMMING:

```
# Install and load required packages
install.packages("tm")
install.packages("class")
library(tm)
library(class)

# Sample dataset (replace with your actual dataset)
political_texts <- c("I support the policies of candidate A.",
                     "I disagree with the actions taken by candidate B.",
                     "The government's decision was commendable.",
                     "I am disappointed with the new legislation.")
sentiments <- c("positive", "negative", "positive", "negative")

# Create a Corpus
corpus <- Corpus(VectorSource(political_texts))

# Data preprocessing
corpus <- tm_map(corpus, tolower)
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, removeWords, stopwords("english"))
corpus <- tm_map(corpus, stripWhitespace)

# Create a Document-Term Matrix
dtm <- DocumentTermMatrix(corpus)

# Convert Document-Term Matrix to a data frame
```

```
dtm_df <- as.data.frame(as.matrix(dtm))
colnames(dtm_df) <- make.names(colnames(dtm_df))

# Splitting the dataset into training and testing sets
set.seed(123)
train_indices <- sample(1:nrow(dtm_df), 0.8 * nrow(dtm_df))
train_data <- dtm_df[train_indices, ]
test_data <- dtm_df[-train_indices, ]

# Model Training
k <- 3 # Number of neighbors
model <- knn(train = train_data, test = test_data, cl = sentiments[train_indices], k = k)

# Model Evaluation
accuracy <- sum(model == sentiments[-train_indices]) / length(sentiments[-train_indices])
print(paste("Accuracy:", accuracy))
```

OUTPUT:

```
[1] Accuracy: 0.5
```

this accuracy is based on a very small and simple dataset. In a real-world scenario, you would need a larger and more diverse dataset for meaningful results.