CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter explains the research methodology used to analyze sentiment related to the "free meal" program promoted by Prabowo and Gibran and public reactions to the program through social media, especially through the X application or Twitter. This methodology includes the process of data collection, data pre-processing, data modeling, to classification using machine learning techniques to identify sentiment patterns (positive, negative, or neutral). This study aims to generate meaningful insights from social media data related to public sentiment towards the program.

3.2 Research Framework

This research framework includes the following steps:

1. Problem Definition and Literature Review
2. Data Collection: Retrieve data from Twitter using specific keywords.
3. Data Pre-processing: Cleaning and preparing data for further analysis.
4. Feature Extraction: Applying stemming and vectorization techniques.
5. Sentiment Classification: Using machine learning models (KNN, Naive Bayes, and SVM).
6. Model Evaluation: Compares model performance using evaluation matrices.

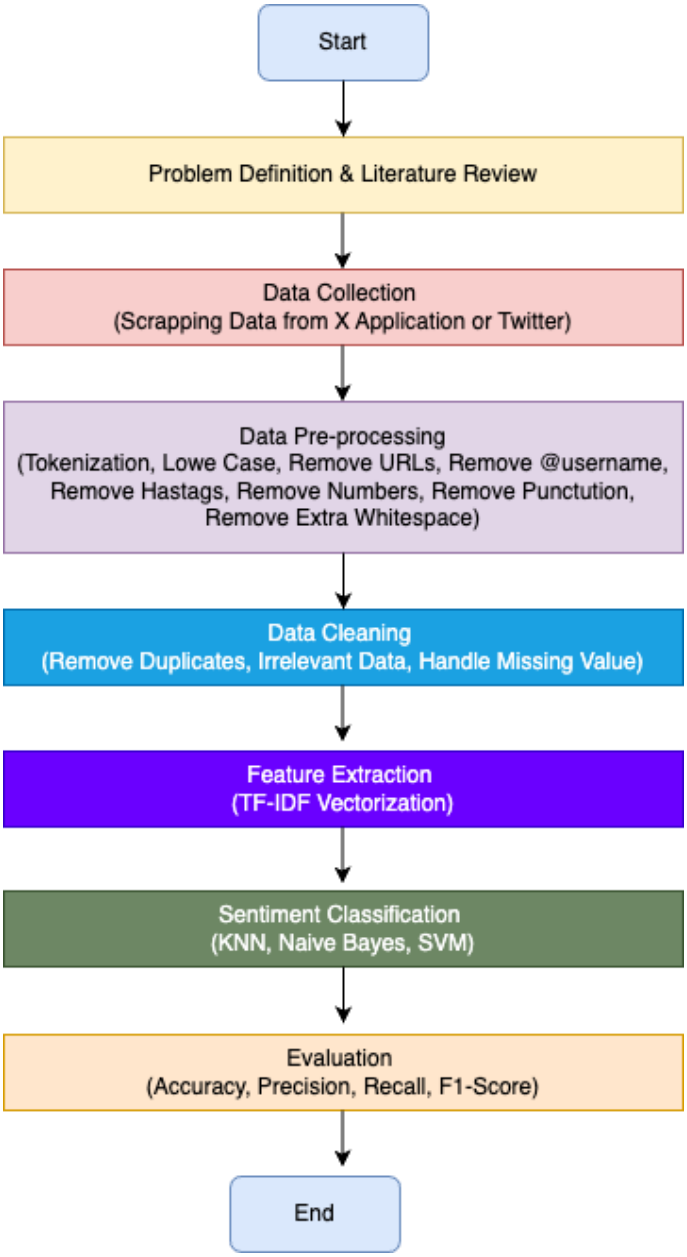


Figure 3.1 General Project Flow

3.3 Problem Formulation

The main objective of this study is to use a sentiment analysis approach to public reactions on social media with machine learning technique classification, thus providing valuable data for further government policies. However, to ensure accurate and reliable analysis, several problems need to be solved.

a. Identifying public sentiment regarding the "free meal" program.

b. Comparing the performance of KNN, Naive Bayes, and SVM algorithms in sentiment classification based on Twitter data.

3.4 Data Collection

Data was collected from the Twitter platform using a web scraping approach. The keywords used for scraping are:

i. "Free meal"
ii. "Free school meal program"
iii. "Prabowo Gibran"

The data collected covers the time span from 2023 to 2024, as well as data prior to 2023 to provide historical context. Information taken includes:

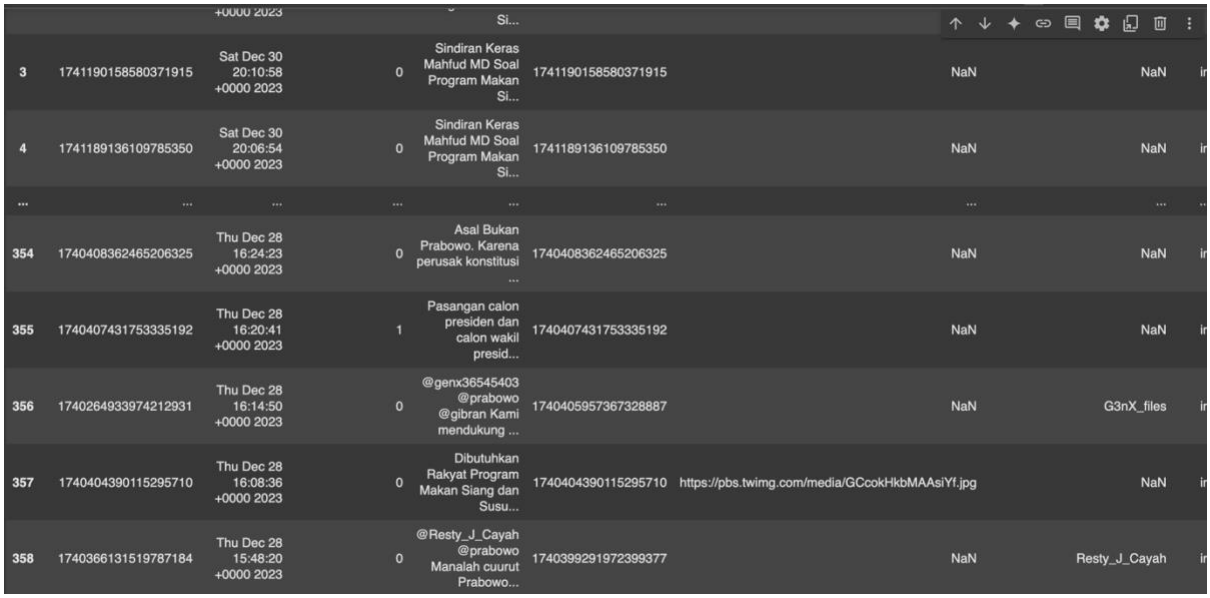
i. Text tweet
ii. Posting date
iii. Username
iv. Number of retweets and likes

For scraping, tools and libraries such as Tweepy and BeautifulSoup are used.

The following dataset was obtained by web scraping process on application X or twitter. The data obtained is related to tweets about Prabowo and Gibran's free meal program. The data is divided into 2 files, namely datasets in 2023 and 2024.

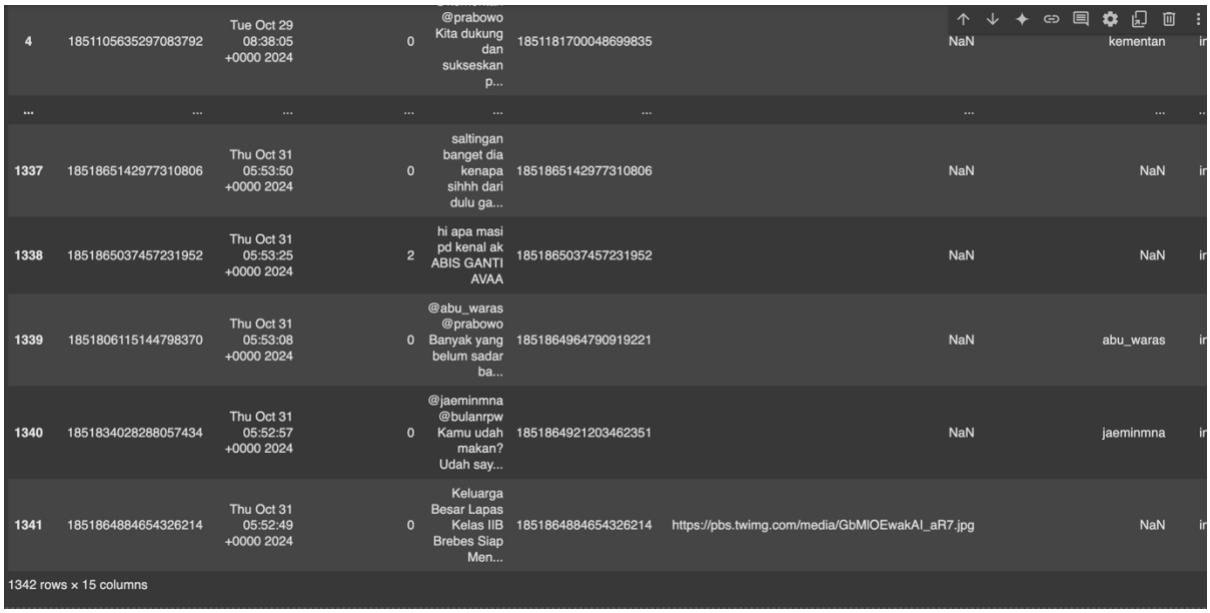
```
# Load and combine datasets
data_2023 = pd.read_csv('/content/dataset_2023.csv')
data_2024 = pd.read_csv('/content/dataset_2024.csv')
data = pd.concat([data_2023, data_2024])
data.dropna()
```

Figure 3.2 Load and Combine Datasets



		+0000 2023		Si...				
3	1741190158580371915	Sat Dec 30 20:10:58 +0000 2023	0	Sindiran Keras Mahfud MD Soal Program Makan Si...	1741190158580371915		NaN	NaN
4	1741189136109785350	Sat Dec 30 20:06:54 +0000 2023	0	Sindiran Keras Mahfud MD Soal Program Makan Si...	1741189136109785350		NaN	NaN
...	...	...	...	...	...	...	...	...
354	1740408362465206325	Thu Dec 28 16:24:23 +0000 2023	0	Asal Bukan Prabowo. Karena perusak konstitusi ...	1740408362465206325		NaN	NaN
355	1740407431753335192	Thu Dec 28 16:20:41 +0000 2023	1	Pasangan calon presiden dan calon wakil presid...	1740407431753335192		NaN	NaN
356	1740264933974212931	Thu Dec 28 16:14:50 +0000 2023	0	@genx36545403 @prabowo @gibran Kami mendukung ...	1740405957367328887		NaN	G3nX_files
357	1740404390115295710	Thu Dec 28 16:08:36 +0000 2023	0	Dibutuhkan Rakyat Program Makan Siang dan Susu...	1740404390115295710	https://pbs.twimg.com/media/GCcokHkbMAAsiYf.jpg	NaN	
358	1740366131519787184	Thu Dec 28 15:48:20 +0000 2023	0	@Resty_J_Cayah @prabowo Manalah cuurut Prabowo...	1740399291972399377		NaN	Resty_J_Cayah

Figure 3.3 Dataset 2023



4	1851105635297083792	Tue Oct 29 08:38:05 +0000 2024	0	@prabowo Kita dukung dan sukseskan p...	1851181700048699835		NaN	kementan
...	...	...	...	...	...	...	...	...
1337	1851865142977310806	Thu Oct 31 05:53:50 +0000 2024	0	saltingan banget dia kenapa sihhh dari dulu ga...	1851865142977310806		NaN	NaN
1338	1851865037457231952	Thu Oct 31 05:53:25 +0000 2024	2	hi apa masi pd kenal ak ABIS GANTI AVAA	1851865037457231952		NaN	NaN
1339	1851806115144798370	Thu Oct 31 05:53:08 +0000 2024	0	@abu_waras @prabowo Banyak yang belum sadar ba...	1851864964790919221		NaN	abu_waras
1340	1851834028288057434	Thu Oct 31 05:52:57 +0000 2024	0	@jaeminmna @bulanrpw Kamu udah makan? Udah say...	1851864921203462351		NaN	jaeminmna
1341	1851864884654326214	Thu Oct 31 05:52:49 +0000 2024	0	Keluarga Besar Lapas Kelas IIB Brebes Siap Men...	1851864884654326214	https://pbs.twimg.com/media/GbMlOEwakAI_aR7.jpg	NaN	

1342 rows × 15 columns

Figure 3.4 Dataset 2024

In the 2023 dataset, 358 data were obtained as shown in Figure 3.1 and in the 2024 dataset, 1342 rows of data were obtained as shown in Figure 3.2. The two datasets are then combined to become a single data unit which will then be processed and analyzed.

3.5 Data Pre-Processing

Initial analysis needs to be completed before moving on to further pre-processing. Data merging procedures are required to unify all the raw data into a single data frame once we have a good understanding of the features available in the data set. Several data processing and data transformation procedures will be used on the data set in an attempt to further unify the disorganized raw data.

Data pre-processing stages include:

1. Tokenization: Breaking text into individual word units or tokens.
2. Lowercase Conversion: Change all text to lowercase to maintain consistency.
3. Stopword Removal: Eliminates common words that do not have significant meaning for sentiment analysis.

Punctuation Removal: Removes punctuation to simplify text.

3.5.1 Initial Analysis

Preliminary analysis is an important step in any data analysis because it helps to become familiar with the data set, understand its structure, format, and the types of variables it contains. Preliminary investigations can identify problems that must be corrected for a reliable analysis, such as missing values, outliers, or contradictions.

In this initial analysis process there are 2 stages that will be carried out, namely:

a. Identify common patterns in raw data.
b. Evaluate data distribution by time and keywords.

3.5.2 Data Cleaning

The data cleaning process is carried out to ensure the quality of the dataset. This step includes:

a. Delete duplicate tweets.

b. Remove tweets that are not in Indonesian or English.

c. Filters ads, spam and irrelevant tweets.

d. Handle missing data through imputation or exclusion.

To identify missing values and remove rows and columns without values, data cleaning is done in this section. Figure 3.5 shows that in data pre-processing, several things are done such as converting text to lower case, removing URLs, removing @username, removing hashtags, removing numbers, removing punctuation and removing extra whitespace. Then apply all the pre-processing processes with the syntax data['full_text'] = data['full_text'].apply(preprocess_text).

```
# Preprocessing function for tweets
def preprocess_text(text):
    text = text.lower()  # Convert text to lowercase
    text = re.sub(r"http\S+|www\S+|https\S+", '', text, flags=re.MULTILINE)  # Remove URLs
    text = re.sub(r'@\w+', '', text)  # Remove @username
    text = re.sub(r'#\w+', '', text)  # Remove hashtags
    text = re.sub(r'\d+', '', text)  # Remove numbers
    text = re.sub(r'[^\w\s]', '', text)  # Remove punctuation
    text = re.sub(r'\s+', ' ', text).strip()  # Remove extra whitespace
    text = stemmer.stem(text)  # Apply stemming
    return text

# Apply preprocessing
data['full_text'] = data['full_text'].apply(preprocess_text)
```

Figure 3.5 Data Cleaning Process

```
import matplotlib.pyplot as plt

# Count the number of duplicates
tweet_bot = len(data.loc[data['full_text'].duplicated() == True])
# Count the number of non-duplicates
tweet_normal = len(data.loc[~data['full_text'].duplicated()])
labels = 'Bot', 'Normal'
sizes = np.array([tweet_bot, tweet_normal])
colors = ['lightskyblue', 'pink']
explode= (0, 0.5)
def absolute_value(val):
    a  = np.round(val/100.*sizes.sum(), 0)

    a=  str(round(val,2))+"%"+'\n'+str(a) +" data"
    return a

plt.pie(sizes, labels=labels, colors=colors,
        autopct=absolute_value, explode=explode,shadow=True)

plt.axis('equal')
plt.title("Data Proportion")
plt.legend()
plt.show()
```

Figure 3.6 Process Cleaning Data and Create Graphs based on Data

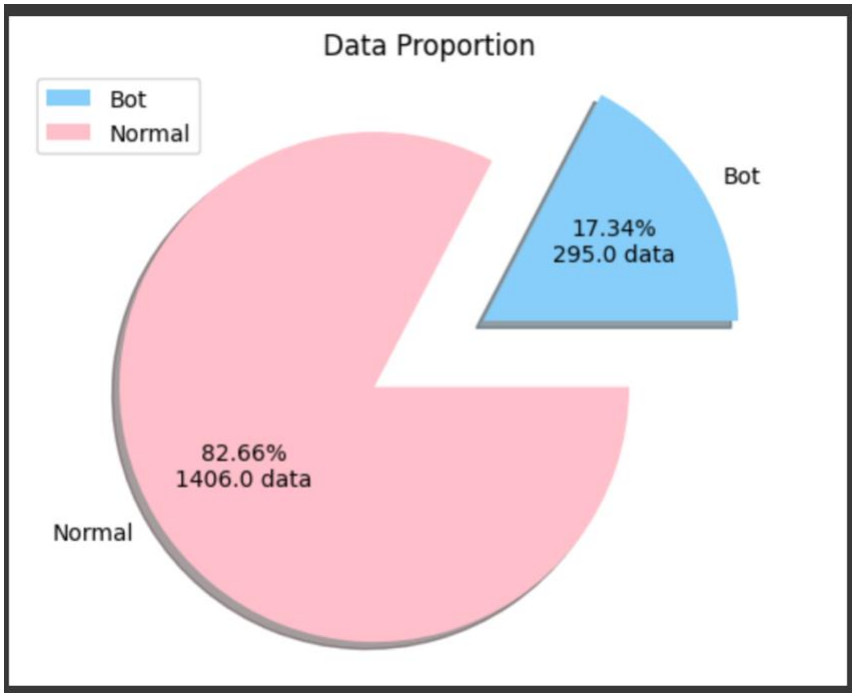


Figure 3.7 Data Proportion

Figure 3.7 shows that from the data that was previously collected from the 2023 and 2024 datasets, a data cleaning process was carried out to obtain a data proportion of which around 82.66% or 1406 data were normal data while 17.34% or 295 data were BOT data.

3.5.3 Data Concatenation

A method of combining data from related datasets into one cohesive dataset in data analysis is called data merging. Since the data was collected in batches, it was essential for this study to merge all datasets from 2023 and 2024. This allowed us to combine all datasets into one for analysis. An example of how we merge data across datasets is shown in Figure 3.8.

```
# Load and combine datasets
data_2023 = pd.read_csv('/content/dataset_2023.csv')
data_2024 = pd.read_csv('/content/dataset_2024.csv')
data = pd.concat([data_2023, data_2024])
data.dropna()
```

Figure 3.8 Load and Combine Dataset

					mahfud ...		
...	...	...	...	...	...	...	
1401	1337	1851865142977310806	Thu Oct 31 05:53:50 +0000 2024	0	saltingan banget dia kenapa sih dari dulu engg...	1851865142977310806	
1402	1338	1851865037457231952	Thu Oct 31 05:53:25 +0000 2024	2	hi apa mas pada kenal aku habis ganti avaa	1851865037457231952	
1403	1339	1851806115144798370	Thu Oct 31 05:53:08 +0000 2024	0	banyak yang belum sadar bahwa presiden saat in...	1851864964790919221	
1404	1340	1851834028288057434	Thu Oct 31 05:52:57 +0000 2024	0	kamu sudah makan sudah sayang aku belum	1851864921203462351	
1405	1341	1851864884654326214	Thu Oct 31 05:52:49 +0000 2024	0	keluarga besar lapas kelas iib brebes siap duk...	1851864884654326214	https://pbs.twimg.com/media/GbMlOEwa

1406 rows × 16 columns

Figure 3.9 Results of the combined data

3.6 Data Modeling

The cleaned data is converted into numerical format using vectorization techniques such as Term Frequency-Inverse Document Frequency (TF-IDF). This representation is used as input for the machine learning model.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, classification_report

# Membagi data menjadi fitur (X) dan target (y)
X = data['full_text']
y = data['sentiment']
```

```
# TF-IDF Vectorizer untuk mengonversi teks ke vektor
vectorizer = TfidfVectorizer(max_features=5000)  # Sesuaikan max_features sesuai kebutuhan
X_vectorized = vectorizer.fit_transform(X)
```

```
# Membagi data menjadi data latih dan uji
X_train, X_test, y_train, y_test = train_test_split(X_vectorized, y, test_size=0.2, random_state=42)
```

Figure 3.10 Process Data Modeling

In Figure 3.10, this is the process of creating a data model. The resulting model will later be entered into the machine learning technique to get the results. The syntax used for the data model creation process is :

vectorizer = TfidfVectorizer(max_features=5000)

X_vectorized = vectorizer.fit_transform(X)

3.7 Stemming Data

Stemming is done to reduce words to their basic form. For example, "eat," "the food," and "ate" all return to "eat." This process helps unite different forms of words that have similar meanings. And in this project we will use the Sastrawi library for the data stemming process.

```
# Initialize Sastrawi stemmer
factory = StemmerFactory()
stemmer = factory.create_stemmer()
```

Figure 3.11 Initialize Sastrawi Stemmer

3.8 Classification Models and Technique

The final stage to obtain sentiment analysis results is to apply and classify the data model into machine learning techniques. The machine learning techniques that will be used are KNN, SVM and Naive Bayes

Three machine learning algorithms are used for sentiment classification:

1. K-Nearest Neighbors (KNN):Classifies tweets based on the majority sentiment of their nearest neighbors in feature space.
2. Naive Bayes: Bayes' theorem based probabilistic model suitable for text classification.
3. Support Vector Machine (SVM): A supervised learning model that separates sentiment classes using hyperplanes in high-dimensional space.

Each model will be evaluated using metrics such as accuracy, precision, recall, and F1-score to determine the best performance.

Model results are evaluated using the following metrics:

a. Accuracy: Percentage of correct predictions.
b. Precision: The accuracy of positive predictions.
c. Recall: The model's ability to detect all positive data.
d. F1-Score: Harmonic mean of precision and recall.

In Figure 3.12 below is the model implementation process in each machine learning technique.

```
# Hyperparameter tuning untuk KNN
knn_params = {'n_neighbors': [3, 5, 7, 9], 'weights': ['uniform', 'distance']}
knn_grid = GridSearchCV(KNeighborsClassifier(), knn_params, cv=5, scoring='accuracy')
knn_grid.fit(X_train, y_train)
knn_best_model = knn_grid.best_estimator_
```

```
# Hyperparameter tuning untuk Naive Bayes
nb_params = {'alpha': [0.1, 0.5, 1.0, 1.5, 2.0]}
nb_grid = GridSearchCV(MultinomialNB(), nb_params, cv=5, scoring='accuracy')
nb_grid.fit(X_train, y_train)
nb_best_model = nb_grid.best_estimator_
```

```
# Hyperparameter tuning untuk SVM
svm_params = {'C': [0.1, 1, 10, 100], 'kernel': ['linear', 'rbf']}
svm_grid = GridSearchCV(SVC(), svm_params, cv=5, scoring='accuracy')
svm_grid.fit(X_train, y_train)
svm_best_model = svm_grid.best_estimator_
```

Figure 3.12 Implementation Model to Machine Learning Technique

3.9 Summary

This chapter explains the research methodology in detail, from data collection to evaluation of the classification model. This process ensures that sentiment analysis of the “free meal” program is conducted systematically and data-driven.