# SENTIMENT ANALYSIS OF SOCIAL MEDIA REACTIONS TO INDONESIA'S FREE MEAL PROGRAM USING MACHINE LEARNING TECHNIQUE

RIZKI SYAPUTRA

UNIVERSITI TEKNOLOGI MALAYSIA

PSZ 19:16 (Pind. 1/13)

**UNIVERSITI TEKNOLOGI MALAYSIA**

**DECLARATION OF THESIS / UNDERGRADUATE PROJECT REPORT AND COPYRIGHT**

Author's full name : RIZKI SYAPUTRA

Date of Birth : 29 AUGUST 1991

Title : SENTIMENT ANALYSIS OF SOCIAL MEDIA REACTIONS TO INDONESIA'S FREE MEAL SCHOOL PROGRAM USING MACHINE LEARNING TECHNIQUE

Academic Session :

I declare that this thesis is classified as:

| | | |
|---|---|---|
| ☐ | **CONFIDENTIAL** | (Contains confidential information under the Official Secret Act 1972)* |
| ☐ | **RESTRICTED** | (Contains restricted information as specified by the organization where research was done)* |
| ✓ | **OPEN ACCESS** | I agree that my thesis to be published as online open access (full text) |

1. I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:
2. The thesis is the property of Universiti Teknologi Malaysia
3. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
4. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

| | |
|---|---|
| **SIGNATURE OF STUDENT** | **SIGNATURE OF SUPERVISOR** |
| MCS211051 | **ASSOC. PROF. DR. MOHD SHAHIZAN BIN OTHMAN** |
| **MATRIX NUMBER** | **NAME OF SUPERVISOR** |
| Date: 16 JANUARY 2025 | Date: 16 JANUARY 2025 |

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

This letter should be written by a supervisor and addressed to Perpustakaan UTM. A copy of this letter should be attached to the thesis.

Date:
Librarian
Jabatan Perpustakaan UTM,
Universiti Teknologi Malaysia,
Johor Bahru, Johor

Sir,

**CLASSIFICATION OF THESIS AS RESTRICTED/CONFIDENTIAL**

**TITLE:** SENTIMENT ANALYSIS OF SOCIAL MEDIA REACTIONS TO INDONESIA'S FREE MEAL PROGRAM USING MACHINE LEARNING TECHNIQUE

**AUTHOR'S FULL NAME:** RIZKI SYAPUTRA

Please be informed that the above-mentioned thesis titled **__________** should be classified as RESTRICTED/CONFIDENTIAL for a period of three (3) years from the date of this letter. The reasons for this classification are

(i)

(ii)

(iii)

Thank you.

Yours sincerely,

**SIGNATURE:**

**NAME:**

**ADDRESS OF SUPERVISOR:** ASSOC. PROF. DR. MOHD SHAHIZAN BIN OTHMAN

"I hereby declare that we have read this thesis and in my
opinion this thesis is suffcient in term of scope and quality for the
award of the degree of Master of Data Science"

| | | |
|---|---|---|
| Signature | : | __________________________________ |
| Name of Supervisor I | : | ASSOC. PROF. DR. MOHD SHAHIZAN BIN OTHMAN |
| Date | : | 17 JANUARY 2025 |

| | | |
|---|---|---|
| Signature | : | __________________________________ |
| Name of Supervisor II | : | |
| Date | : | |

| | | |
|---|---|---|
| Signature | : | __________________________________ |
| Name of Supervisor II | : | |
| Date | : | |

**Declaration of Cooperation**

This is to confirm that this research has been conducted through a collaboration Rizki Syaputra and University Technologi Malaysia (UTM)

Certified by:

Signature :

Name :

Position :

Official Stamp

Date

* This section is to be filled up for theses with industrial collaboration

**Pengesahan Peperiksaan**

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar :

Nama dan Alamat Pemeriksa Dalam :

Nama Penyelia Lain (jika ada) :

**Disahkan oleh Timbalan Pendaftar di Fakulti:**

Tandatangan :

Nama :

Tarikh :

# SENTIMENT ANALYSIS OF SOCIAL MEDIA REACTIONS TO INDONESIA'S FREE MEAL PROGRAM USING MACHINE LEARNING TECHNIQUE

RIZKI SYAPUTRA

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Science (Data Science)

School of Computing
Faculty of Computing
Universiti Teknologi Malaysia

JANUARY 2025

# DECLARATION

I declare that this thesis entitled *"Sentiment Analysis Of Social Media Reactions To Indonesia's Free Meal Program Using Machine Learning Technique"* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

| | | |
|---|---|---|
| Signature | : | .................................................... |
| Name | : | RIZKI SYAPUTRA |
| Date | : | 17 JANUARY 2025 |

## DEDICATION

This thesis is dedicated to my mother and father, the most important people in my life. Mother and father, although you have never received higher education, cannot read and write. But look, now your child has reached the stage of master's degree. I hope I can make you proud to have a child like me. I love you, father and mother.

## ACKNOWLEDGEMENT

In preparing this thesis, I have been in contact with many people, researchers, academics, and practitioners. They have contributed to my understanding and thinking. In particular, I would like to express my sincere appreciation to my main thesis supervisor, ASSOC. PROF. DR. MOHD SHAHIZAN BIN OTHMAN, for his encouragement, guidance, criticism, and friendship.

To my father and mother, my greatest motivation to be able to complete this master's degree. I want to make you proud. Although from a simple family from a village with a house at the foot of a hill, but look now your child has been able to enjoy a master's degree. And to my small family, my wife, and my two children. You are my greatest dream to be able to bring you to a better stage of life.

My fellow postgraduate students must also be acknowledged for their support. My sincere appreciation also goes to all my colleagues and others who have provided assistance on various occasions. Their views and suggestions are indeed useful. Unfortunately, it is impossible to list them all in this limited space.

# ABSTRACT

The purpose of this study is to  analysis public sentiment towards the free school meal program organized by the Indonesian Government using data from social media. This program aims to improve students' nutritional intake and help ease the burden on low-income families. However, this policy has drawn mixed reactions on social media, both in the form of support and criticism. Data was collected from various social media platforms and processed through data cleaning and modelling stages using machine learning algorithms, namely Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Naive Bayes. The results of the analysis show that the SVM and Naive Bayes methods have a higher level of accuracy in classifying positive and negative sentiments than KNN. Overall, the majority of public sentiment towards this program is positive, reflecting support for the policy that is considered beneficial for students in need. However, there are also a number of negative sentiments that highlight issues of distribution and quality of service. These findings can provide insights for the government to evaluate and improve the effectiveness of the program, as well as identify areas that need improvement.

# ABSTRAK

Tujuan kajian ini adalah untuk menganalisis sentimen masyarakat terhadap program makan sekolah percuma anjuran Kerajaan Indonesia menggunakan data dari media sosial. Program ini bertujuan untuk meningkatkan pengambilan nutrisi pelajar dan membantu meringankan beban keluarga berpendapatan rendah. Bagaimanapun, dasar ini telah mendapat pelbagai reaksi di media sosial, baik dalam bentuk sokongan mahupun kritikan. Data dikumpul daripada pelbagai platform media sosial dan diproses melalui peringkat pembersihan dan pemodelan data menggunakan algoritma pembelajaran mesin, iaitu Mesin Vektor Sokongan (SVM), K-Nearest Neighbors (KNN), dan Naive Bayes. Hasil analisis menunjukkan kaedah SVM dan Naive Bayes mempunyai tahap ketepatan yang lebih tinggi dalam mengklasifikasikan sentimen positif dan negatif berbanding KNN. Secara keseluruhannya, majoriti sentimen masyarakat terhadap program ini adalah positif, mencerminkan sokongan terhadap dasar yang dianggap bermanfaat untuk pelajar yang memerlukan. Walau bagaimanapun, terdapat juga beberapa sentimen negatif yang menonjolkan isu pengagihan dan kualiti perkhidmatan. Penemuan ini boleh memberi pandangan kepada kerajaan untuk menilai dan menambah baik keberkesanan program, serta mengenal pasti bidang yang memerlukan penambahbaikan.

# TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| ANN | - | Artificial Neural Network |
| GA | - | Genetic Algorithm |
| PSO | - | Particle Swarm Optimization |
| MTS | - | Mahalanobis Taguchi System |
| MD | - | Mahalanobis Distance |
| TM | - | Taguchi Method |
| UTM | - | Universiti Teknologi Malaysia |
| XML | - | Extensible Markup Language |
| ANN | - | Artificial Neural Network |
| GA | - | Genetic Algorithm |
| PSO | - | Particle Swarm Optimization |

# LIST OF SYMBOLS

| | | |
|---|---|---|
| $\delta$ | - | Minimal error |
| $D, d$ | - | Diameter |
| $F$ | - | Force |
| $v$ | - | Velocity |
| $p$ | - | Pressure |
| $I$ | - | Moment of Inersia |
| $r$ | - | Radius |
| $\mathrm{Re}$ | - | Reynold Number |

# LIST OF APPENDICES

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

Social media has grown in popularity and effectiveness in recent years as a means of rapidly disseminating information. A large number of individuals utilize it effectively, particularly in Indonesia. A lot of individuals oppose government initiatives on social media. The Free Meal Program has been the most talked about program and a trending issue on different social media platforms since the presidential election debate, the presidential election process, and the inauguration of Gibran Rakamubi Raka and Mr. Prabowo Subianto as Indonesia's eighth president in 2024. There have been mixed reactions to this program on social media; some have been complimentary, while others have been hostile or cynical. The Python programming language, Google Collab, and machine learning are used in this work to apply sentiment analysis approaches. Finding trends in the public's perception of this program is the aim of this study. The goal of this study is to shed light on the efficacy of the free meal program so that it may serve as a foundation for future program evaluations and improved laws.

## 1.2 Problem Background

The Indonesian government introduced the Free Meals Program as a way to improve the nutritional status of the poor and help communities cope with the economic crisis, but its success has been largely based on the acceptance of the Indonesian public. Although complex and unstructured, social media offers a wealth of information about public reactions in real time. To overcome this difficulty and understand how the public perceives the Prabowo-Gibran administration's free meal program, an appropriate data analysis method is needed, which involves applying sentiment analysis to Twitter data. This knowledge is essential for policymakers to make proper assessments of the program.

## 1.3 Problem Statement

Social media has become one of the main platforms for people to express their opinions on government policies, including social programs such as the Free Meal Program. As one of the Indonesian government's efforts to improve the nutritional intake and welfare of students from low-income families, this program has great potential to provide social benefits. However, public responses to this program are not always uniform.

A variety of sentiments, both positive and negative, appear on social media, reflecting support, criticism, or dissatisfaction. Negative sentiments often highlight issues such as uneven distribution of aid, questionable food quality, or lack of transparency in program implementation. On the other hand, positive sentiments indicate that many people appreciate this government initiative as a strategic step to reduce the economic burden on the underprivileged.

Unfortunately, this broad public response has not been systematically analyzed to provide reliable insights to policymakers. Without a deep understanding of public perception, the government risks facing challenges in increasing public acceptance and program effectiveness. Therefore, machine learning-based sentiment analysis is needed to identify sentiment patterns on social media accurately and efficiently. Thus, the results of the analysis can be the basis for continuous evaluation and improvement of policies.

This issue is relevant to ensure that public policies such as the Free Meal Program are not only implemented well, but also receive positive acceptance from the wider community.

## 1.4 Research Question

The research question of the research are as follows:

a. What is the general sentiment of the public on Social Media towards the free meal program launched by the Prabowo-Gibran government?

b. Is there a certain pattern in sentiment based on time or discussion theme?

c. What factors influence positive and negative sentiment regarding this program?

## 1.5 Research Aim

This project aims to analyze public reactions on social media to the Prabowo-Gibran government's free meal program using sentiment analysis to understand public opinion patterns and factors that influence positive and negative perceptions.

## 1.6 Research Objectives

The research objectives of this research are follows:

To Gather Twitter information on the free meal program using crawling methods.

a. To Implement sentiment analysis to the gathered data in order to determine the public's perspective (positive, negative, or neutral).

b. To Categorize the sentiment-related primary subjects of public discourse.

c. To Present research to help policymakers and the government improve the caliber of social welfare initiatives.

## 1.7 Research Scope (Current Work)

As with all other researchers, the production and assessment of this study are subject to the following boundaries and limitations:

a. Data Collection: The Prabowo-Gibran government's free meal program is the source of data used, which was collected via Twitter using certain keywords and hashtags.

b. Sentiment Analysis Method: Naive Bayes Classifier, SVM, and KNN machine learning techniques were used in the analysis process.

c. Topic Modelling: To find important themes in the data, unsupervised learning techniques will be applied.

d. Time Frame: To understand the immediate response of the public, data analysis is limited to a specific time frame around the launch of the program. The data used is relevant to this program and comes from 2023 until 2025.

## 1.8 Expected Research Contribution

This research is expected to provide an in-depth understanding of public perceptions of social policies through social media data analysis. The findings of this study can be valuable input for the government to evaluate and improve welfare programs in the future. In addition, this study will provide examples of the implementation of sentiment analysis that can be applied to other policies.

## 1.9 Thesis Organization

The remaining sections of the thesis are structured as follows:

Chapter 2 provides an extensive review of the literature on, Sentiment Analysis and Social Media, Free Meal Program in Indonesia, Sentiment Analysis Techniques for Public Policy, and classification techniques. It encompasses the research background, explores existing research gaps, and delves into the current state of the study.

Chapter 3 shows the direction of the proposed research methodology for this study. Next, the method sentiment analysis techniques  is presented.

Chapter 4 describes the proposed method's findings and expected results for sentiment analysis for this research.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

This chapter reviews existing literature and explores academic research issues, highlighting research issues within the broad scope of global scientific understanding. The chapter begins with a brief introduction to free meal school program, sentiment analysis data, and classification technique, as well as a brief overview of machine learning for classification, which provides a foundation for understanding the research effort.

## 2.2 Free Meal School Program

The Free Meal Program or “Free Nutritious Meals” (MBG) initiated by the Prabowo-Gibran pair has officially been discussed in the 2025 budget planning. The Ministry of Finance (Kemenkeu), the Coordinating Ministry for Economic Affairs (Kemenko) and the Prabowo-Gibran transition team have set the MBG budget at IDR 71 trillion in the first phase in 2025. This amount is considered to have taken into account the fiscal deficit target of 2.29% - 2.82% (Prabowo-Gibran, 2023; BBC Indonesia, 2024).

The MBG program in the first phase will focus on targeting elementary, junior high and senior high school students in the quintile 1 and 2 categories in the disadvantaged, outermost and remote areas (3T) in Indonesia (BBC Indonesia, 2024). According to the Prabowo-Gibran transition team, the target targets, budget size and program governance will continue to be evaluated and expanded to eradicate stunting in Indonesia.

The amount of budget to be spent and the effectiveness of the impact that will be generated from this program become the pros and cons of public discourse.

Moreover, public budgets and public policies should be accounted for by policy makers. In addition, changes related to program names, targets, budgets and so on are known to civil society only through media coverage. There are no permanent, transparent and sustainable channels and mechanisms for public participation to ensure civil society participation in monitoring program developments. Various public concerns have emerged; from the quality of planning, limited fiscal space, to unclear governance (Suwastoyo, 2024). Given the urgency, the Center for Indonesia's Strategic Development Initiatives (CISDI) took the initiative to conduct a study of the MBG program which is divided into several series. The state budget is expected to be used transparently, measurably and have a positive impact on public health development.

Overall, the free lunch program initiated by Prabowo Subianto and Gibran Rakabuming Raka is an initiative that has the potential to have a significant impact on the Indonesian people. By promising and implementing this program, the presidential and vice presidential candidate pair number 2 have shown their commitment to improving the social and economic welfare of the people.

From a social perspective, this program has a positive impact in reducing hunger levels, increasing access to education, and empowering local communities through the participation of BUMDes, UMKM, and cooperatives in the food supply chain. In addition, this program also provides real assistance to families in difficult economic conditions by reducing their economic burden and increasing access to adequate food. From an economic perspective, this program provides a significant stimulus to the food sector, food processing industry, and distribution sector, which results in local and national economic growth.

The provision of free meals in schools also helps increase consumption of goods and services, creates new business opportunities, and supports the growth of small and medium enterprises. However, to ensure the sustainability and effectiveness of this program, careful evaluation is needed. This evaluation must take into account various aspects, from the quality of the food provided to the efficiency of resource use. The implications of this evaluation will influence the direction and strategy of further program development, as well as ensuring that the program continues to provide maximum benefits to the Indonesian people as a whole. Thus, the free lunch program is not only a solution to the problem of hunger, but also a progressive step in building inclusive and sustainable social and economic welfare in Indonesia.

With a strong commitment and good cooperation between the government, community, and private sector, this program has the potential to provide a positive impact on the future of the nation and state.

Twitter or currently known as the X application is a social media platform that allows its users to interact through what is called a tweet. Users can share information and views on various topics being discussed. X has features that allow unlimited submission of opinions, search for the latest news, share other people's tweets, and provide comments. In the X application, information can spread quickly and easily, making it a means of finding out someone's opinion sentiment, both positive and negative.

Sentiment analysis is a method for assessing the emotional tone in digital text to determine whether the tone is positive, negative, or neutral. The use of sentiment analysis involves analyzing opinions, feelings, evaluations, emotions, assessments, or attitudes towards a product, figure, organization, issue, service, and event. In addition, sentiment analysis is always related to society because the information obtained comes from social media where society acts as its users.

## 2.3 Social Media

Social media is a digital platform that is very influential in today's human life. Its main purpose is to facilitate communication between people, share information, and create social interactions without any limitations of space and time. Social media is an online media tool that functions with web-based technology and changes the way of communicating from one way to a dialogic or two-way interaction.

In this digital era, public opinion is often expressed through social media, one of which is Platform X. This platform is the main place for the public to convey views, comments, and criticisms of current issues. The Free Lunch Program policy, for example, has received great attention on Platform X. The many opinions circulating on social media regarding this program show the importance of analyzing public responses. This method allows researchers to automatically identify and classify public opinion based on text generated by social media users. (Program et al., 2024)

## 2.4 Sentiment Analysis

### 2.4.1 Definition and Importance of Sentiment Analysis

Sentiment analysis is the process of utilizing text analytics to gather information from various sources on the internet. There are several types of sentiment analysis, such as fine-grained sentiment analysis that focuses on the polarity of opinions, intent sentiment analysis that attempts to find the motivation behind a user's message, and aspect-based sentiment analysis that associates a particular sentiment with an aspect of a product or service.

Sentiment analysis is the process of analysing digital text to determine whether the emotional tone of the message is positive, negative, or neutral.(Awazon, 2022)

The goal of sentiment analysis is to evaluate the emotions, attitudes, and opinions expressed by individuals across platforms regarding goods, brands, services, politics, or organizations. This approach includes machine learning and is lexicon-based, which allows grouping data into categories such as very positive, positive, neutral, negative, and very negative.

Sentiment analysis consists of various types such as emotion detection, aspect-based sentiment analysis, and fine-grained sentiment analysis. Fine-grained sentiment analysis is a type of analysis that provides specific assessments and is commonly applied in the field of e-commerce. Emoticon detection is a type of analysis that aims to identify emotions in messages, such as happy, sad, angry, and others. Aspect-based sentiment analysis is a type of analysis that aims to identify influential aspects and assessments from customers.

### 2.4.2 The Relevance of Sentiment Analysis to Social Media

Social media has become a very important source of data in sentiment analysis. Platforms such as Twitter, Facebook, and Instagram provide a space for people to express their views on a variety of topics, including public policy. Collecting and analyzing information expressed by audiences on social media allows for a deeper understanding of the attitudes, opinions, and emotions underlying the text. (Ratna Patria, 2022). In the context of public policy, such as the free school meal program in Indonesia, sentiment analysis can provide insight into the level of acceptance, criticism, and public perception of the program.

## 2.5 Machine Learning Technique for Sentiment Analysis

### 2.5.1 Various Machine Learning Algorithm

Various machine learning algorithms have been used for sentiment analysis, each with its strengths and limitations:

a. Naive Bayes: A probabilistic classifier based on Bayes' theorem, which performs well in text classification due to its simplicity and efficiency.

b. Support Vector Machines: A supervised learning model that finds the optimal hyperplane to separate classes in a feature space, robust in high-dimensional spaces.

c. Random Forest: This is an ensemble learning methodology where multiple decision trees are created and then their results are combined to perform better for classification tasks without overfitting.

d. Neural Networks: These models take inspiration from the way the human brain works; thus, neural networks are capable of learning very complex patterns in data. Variants of deep learning such as Convolutional Neural Networks and

Long Short-Term Memory networks have performed impressively on sentiment analysis tasks.

A comparative study by (Kolchyna et al., 2015) demonstrated the superiority of machine learning methods, especially SVM and Naive Bayes classifiers over lexicon-based approaches when performing sentiment classification tasks.

### 2.5.2 Naïve Bayes

The reason Naïve Bayes classifier can be one of the most popular algorithms used for text mining is that it is convenient to use. Besides, it has a fast processing time, is easy to implement with a reasonably simple structure, and it is highly effective. MNB calculates the probability of a class mainly according to its attributes and defines the class that has the highest probability. It categorizes classes mainly using simple opportunities by assuming that every attribute in the data is mutually unique. (Ajhari, 2023)In the opportunity model, every k class and the number of attributes can be written as in (1).

$$P = (Y1|x1, x2, \dots, Xn) \tag{1}$$

The probability of occurrence of feature data Xa in category class YkP(xa|yk) is the result of MNB calculation multiplied by the probability of category class P(yk). The distribution of feature data P(xa) will occur from the results of the previous calculation. Thus, the new calculation is determined in equation (2).

$$P(y_k|x_a) = \frac{P(y_k)P(x_a|y_k)}{P(x_a)} \tag{2}$$

Then, the highest probability value is selected from each probability class to determine the optimal class. In formula (3) is the selection of the highest value.

$$y(x_i) = \arg\max P(y) \prod_{i=1}^{a} P(x_1|y) \tag{3}$$

The accuracy of classification does not only depend on probability but can also use weights in each class. In this way, attributes can increase the predictive effect.

### 2.5.3 Support Vector Machine

According to (Ajhari, 2023), support vector machine (SVM) has a solution, namely the maximum margin classifier or hyper-plane concept can overcome the prediction classification problem that occurs in other linear classifiers. In classifying correctly It is known from the mathematical calculation of the SVM method prediction.

Solving for the enlargement and reduction of two distances can allow the margin to be maximized in equation 4, thus allowing equation 5 to minimize and find the sum of the distances from the separating hyper-plane to the nearest point.

$$Margin = \frac{2}{||\vec{w}||^2} \tag{4}$$

$$L(w) = \frac{\overrightarrow{||w||}^2}{2} \tag{5}$$

The emergence of finite optimization problems can be done using an approach using numerical calculations in (6) below.

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 \end{cases} \tag{6}$$

### 2.5.4 K-Nearest Neighbors

K-Nearest Neighbors is a classic supervised machine learning algorithm used in various classification and regression tasks. This algorithm follows the intuitive idea that similar data tend to cluster close to each other in the feature space. KNN predicts the output based on finding the 'k'-Neighbors-most similar data points to the given input based on the majority class for classification or the mean value for regression. It is a nonparametric method, hence does not assume any fundamental data distribution, hence finds applications in various fields.

KNN works by storing all available examples and then classifying new examples based on some similarity measure, such as Euclidean distance. The algorithm works as follows:

a. Store all the data in memory.

b. Compute the distance between the new input and all stored examples.

c. Select a number of 'k' examples that have the smallest distance.

Predict the output by taking the majority class among the neighbors in the case of classification or calculating the average in the case of regression.

The performance of the algorithm is greatly affected by the choice of 'k'. If 'k' is small, then the algorithm is sensitive to noise; if 'k' is large, it may smooth out the class boundaries. Thus, the optimal 'k' is important and is usually determined by cross-validation.

## 2.6 Previous Work on Sentiment Analysis

Sentiment analysis is increasingly used in public policy assessments to gauge citizen reactions. For example, studies on health policies during COVID-19 have highlighted important insights into public concerns and levels of support (Wang et al., 2021). In the education space, sentiment analysis has been used to assess programs such as school meal initiatives, offering data-driven insights for program optimization. A study by Kumar et al. in 2023 found evidence of the effectiveness of sentiment analysis in evaluating policies, integrating social media data to understand public satisfaction and areas for improvement.

Analyzing sentiments on social media, however, comes with its specific challenges. There is noisy data filled with slang, abbreviations, and grammatically incorrect sentences, complicating text processing. The informal and code-mixed varieties of Indonesian languages also add to the intricacies of sentiment analysis. These challenges would need specific preprocessing techniques and a strong linguistic model. Works like Rahimi et al. in 2020 discuss the challenges of analyzing informal language texts in low-resource settings. Table 2.1 summarizes previous work on sentiment analysis leveraging social media and public policy with several machine learning techniques.

Table 2. 1 Previous work on Sentiment Analysis

| **Author /Year** | **Title** | **Research Focus** | **Machine Learning Methods** |
|---|---|---|---|
| (Monselise et al., 2021) | Topics and sentiments of public concerns regarding COVID-19 vaccines: Social media trend analysis | Analysis of social media trends related to public sentiment towards vaccination | NLP, Sentiment Classification |
| (Adak et al., 2022) | Sentiment analysis of customer reviews of food delivery services using deep learning and explainable AI | Food delivery service customer review analysis using explainable AI | Deep Learning |
| (Ainin et al., 2020) | Sentiment analysis of multilingual tweets on halal tourism | Sentiment on halal tourism through multilingual tweet analysis | Random Forest, Naive Bayes |
| (Nguyen et al., 2019) | Pride, love, and twitter rants: Combining machine learning and qualitative techniques | Combination of machine learning and qualitative techniques for social media sentiment analysis | Sentiment Analysis with qualitative features |

| | | | |
|---|---|---|---|
| (Hudaefi et al., 2022) | Zakat administration in times of COVID-19 pandemic in Indonesia: A knowledge discovery via text mining | Text mining on zakat administration during the pandemic | Text Mining |
| (Tao et al., 2019) | Social media data-based sentiment analysis of tourists' air quality perceptions | Sentiment analysis of air quality perception by tourists | Neural Networks |
| (Landwehr et al., 2016) | Using tweets to support disaster planning, warning and response | Tweet analysis to support disaster planning | Logistic Regression, Support Vector Machine |
| (Sudo et al., 2020) | Robots, AI, and service automation (RAISA) in hospitality: Sentiment analysis of YouTube streaming data | Sentiment analysis of YouTube streaming data about RAISA | BERT, Contextual Analysis |

## 2.7 Research Gap

Sentiment analysis has great potential for evaluating public policies, there are several research gaps related to free meal programs in Indonesia. Most existing studies tend to focus on evaluating general policies or other education programs without specific attention to free meal programs. Relevant studies often highlight global or regional policies, such as poverty reduction and school nutrition, but few have explored the impact of these programs directly through sentiment analysis on social media.

## 2.8 Summary

This chapter includes a literature review of ongoing research regarding classification and sentiment analysis using social media. This chapter presents an analysis of the similarities and differences between various methods, policies, and algorithms. Apart from that, this chapter also provides an in-depth discussion regarding the Free Meal Program from Indonesia's Government. The next chapter will discuss the research methodology and outline the main strategies used in this thesis.

# 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. Also 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 modelling, 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.

The details of the research framework for this study are shown in Figure 3.1.

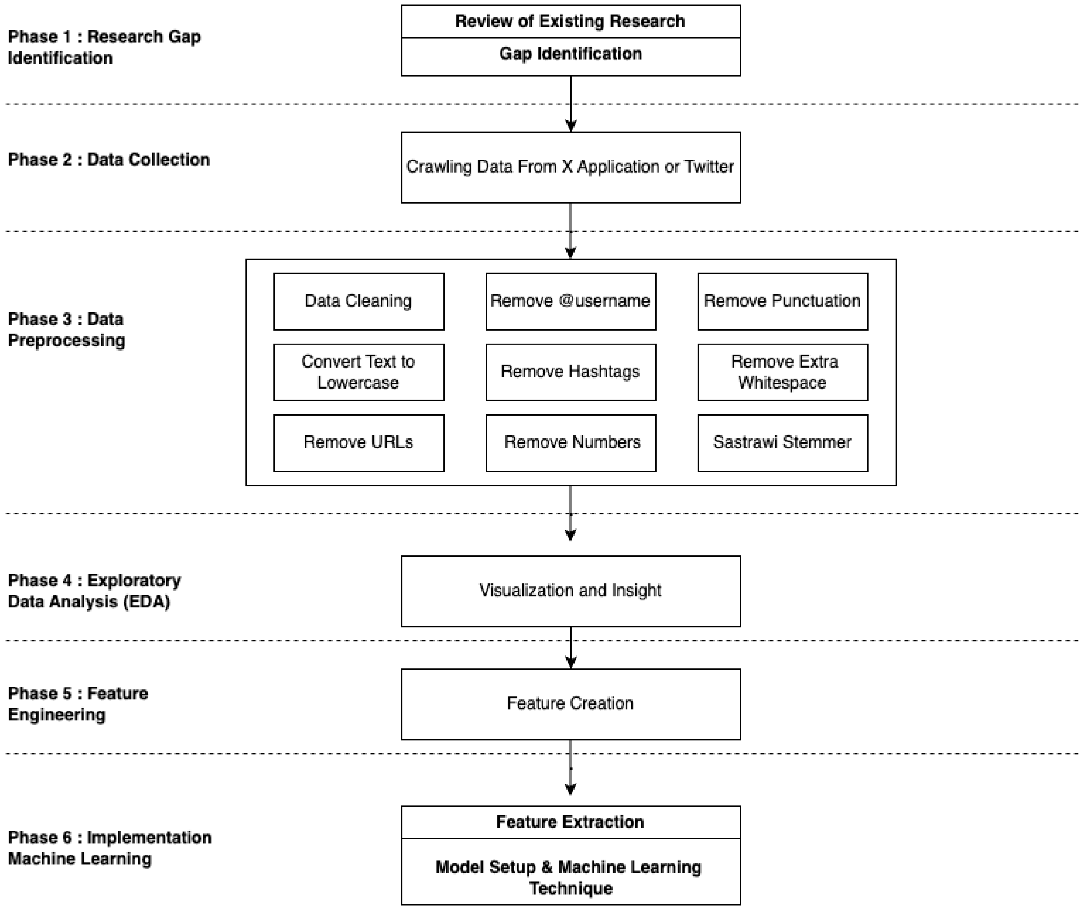



Figure 3. 1 Research Framework For Sentiment Analysis

## 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 Crawling Data Technique. The keywords used for crawling data are:

a) "Free meal"
b) "Free school meal program"
c) "Prabowo Gibran"

The data collected covers the time span from 2023 to 2025, as well as data prior to 2023 to provide historical context. Information taken includes:

a) Text tweet
b) Posting date
c) Username
d) Number of retweets and likes

The following dataset was obtained by crawling data process on application X or twitter. The data obtained is related to tweets about Prabowo and Gibran's free meal program. This data collected from 2023 until January 2025.

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
import pandas as pd
import numpy as np
import re
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

data = pd.read_csv('/content/drive/My Drive/content/dataset/final_dataset.csv')
data
```

Figure 3. 2 Load The Dataset

The total dataset that has been collected is 2916 rows of data with 15 columns as shown in the Figure 3.3

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **2** | 1866464297657975123 | Wed Dec 11 21:21:34 +0000 2024 | 0 | @StevanFirman15 @prabowo Yakin pak makan grati... | 1866956515414094156 | | |
| **3** | 1866754957552259364 | Wed Dec 11 18:42:05 +0000 2024 | 0 | @03__nakula Makanya program Prabowo makan sian... | 1866916381872230896 | NaN | 03__nakula |
| **4** | 1866877557489926164 | Wed Dec 11 16:07:49 +0000 2024 | 0 | Prancis Dukung Program Makan Bergizi Gratis Pr... | 1866877557489926164 | NaN | NaN |
| ... | ... | ... | ... | ... | ... | ... | ... |
| **2911** | 1851865142977310806 | Thu Oct 31 05:53:50 +0000 2024 | 0 | saltingan banget dia kenapa sihhh dari dulu ga... | 1851865142977310806 | NaN | NaN |
| **2912** | 1851865037457231952 | Thu Oct 31 05:53:25 +0000 2024 | 2 | hi apa masi pd kenal ak ABIS GANTI AVAA | 1851865037457231952 | NaN | NaN |
| **2913** | 1851806115144798370 | Thu Oct 31 05:53:08 +0000 2024 | 0 | @abu_waras @prabowo Banyak yang belum sadar ba... | 1851864964790919221 | NaN | abu_waras |
| **2914** | 1851834028288057434 | Thu Oct 31 05:52:57 +0000 2024 | 0 | @jaeminmna @bulanrpw Kamu udah makan? Udah say... | 1851864921203462351 | NaN | jaeminmna |
| **2915** | 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 |

2916 rows × 15 columns

Figure 3. 3 The Dataset Preview

## 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. Table 3.1 lists every detail of the data pre-processing that was used.

Table 3. 1 Data Pre-processing Methods

| **Data Pre-Processing** | **Purpose** |
|---|---|
| Preliminary Analysis | To evaluate the provided dataset and obtain insightful knowledge for the modelling phase that follows |
| Data Cleaning | Find the missing value and eliminate the rows that do not have it |
| Data Visualization | A pie chart illustrating the trend of each variable for sentiment analysis free meal program. |

### 3.5.1 Preliminary 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

Data cleaning is an important process in sentiment analysis, especially to ensure that the data used is clean, relevant, and can be processed well by the model. Here are the data cleaning steps carried out on the Twitter tweet dataset about the Prabowo-Gibran free meal program:

1. **Initialize Sastrawi Stemmer**

   Sastrawi is used to perform stemming, which is changing affixed words into basic forms (root words). For example, "`makannya`" becomes "`makan`". Using Stemming helps simplify word variations so that the model can more easily recognize patterns in the data.

2. **Convert Text to Lowercase**

   All letters in the text are converted to lowercase. Makes the analysis more consistent because uppercase and lowercase are treated the same. For example, "`PRABOWO`" and "`prabowo`" are considered the same.

3. **Remove URLs**

   Removes links (URLs) from text such as "https://...". Links do not provide relevant sentiment information and can interfere with analysis.

4. **Remove @username**

   Removes mentions or tags such as "@user". Mentions are usually not relevant for sentiment analysis because they only point to a specific account.

5. **Remove Hashtags**

   Removing hashtags such as "`#prabowo`" or "`#makangratis`". Hashtags can be removed because they often do not contain the context needed in sentiment analysis, although there are certain cases where hashtags are analyzed separately.

6. **Remove Numbers**

   Removes numbers from text. Numbers usually have no meaning in the context of sentiment, unless specifically relevant (can be processed separately if important).

7. **Remove Punctuation**

   Removes punctuation such as ".", ",", "?", etc. Punctuation does not contribute directly to sentiment analysis.

8. **Remove Extra Whitespace**

   Removes excess whitespace in text. Makes text neater and easier to read.

9. **Apply Stemming**

   Uses the Sastrawi stemmer to convert words to their basic form. Reduces variations in words that have the same meaning.

10. **Apply Preprocessing**

    Combines all the above steps into one preprocessing pipeline that is applied to the entire dataset. Ensures all data is processed in a uniform manner.

11. **Translate Data to Minimize English Words**

Translates English words to Indonesian using a library such as the Google Translate API. Standardizes the language so that all text is in one language (Bahasa Indonesia) to facilitate sentiment analysis.

In the Figure 3.4, it explains the flow of the data cleaning process with a literary stemmer to the process of minimizing words in English.

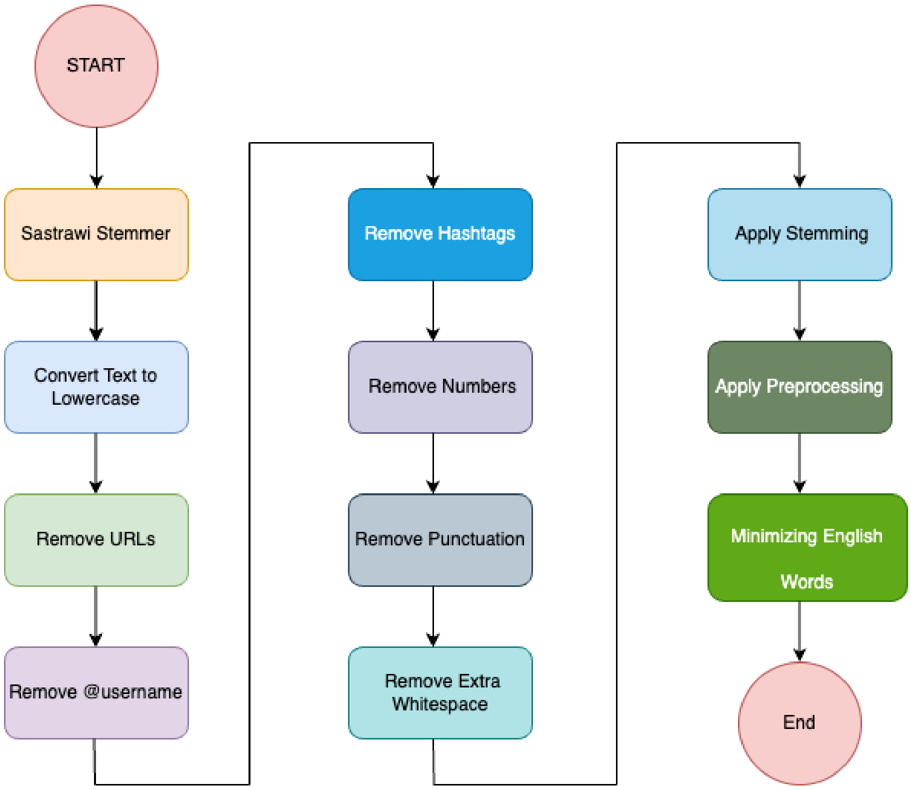



Figure 3. 4 Flow Data Cleaning and Preparation

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)`.

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

# 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

```python
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

Figure 3.7 shows that from the data that was previously collected from the 2023 – January 2025 datasets, a data cleaning process was carried out to obtain a data proportion of which around 85.91% or 2505 data were normal data while 14.09% or 411 data were BOT data.

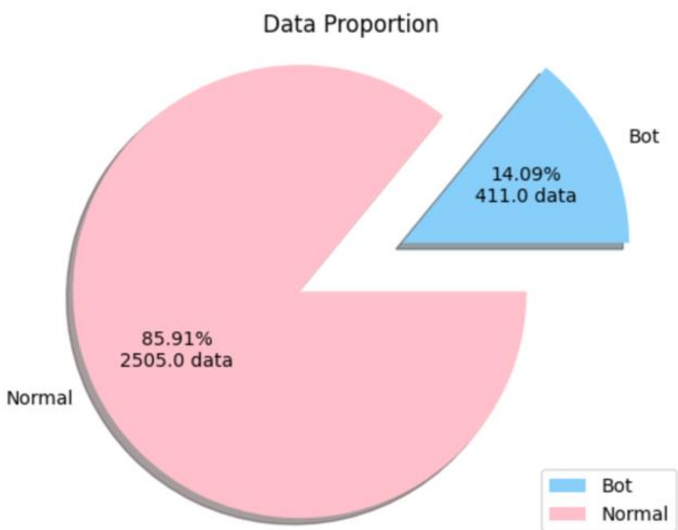



Figure 3. 7 Data Proportion

## 3.6 Data Modelling

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.

In Figure 3.8, this is the process of creating a data model. The resulting model will 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)
```

```
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

# Divide the data into features (X) and targets (y)
X = data['full_text']
y = data['sentiment']
```

```
# TF-IDF Vectorizer to convert text to vectors
vectorizer = TfidfVectorizer(max_features=5000)  # Adjust max_features as needed
X_vectorized = vectorizer.fit_transform(X)
```

```
# Splitting data into training and test data
X_train, X_test, y_train, y_test = train_test_split(X_vectorized, y, test_size=0.2, random_state=42)
```

Figure 3. 8 Process Data Modelling

### 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. 9 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 3.10 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_
```

```
# Predicting test data and calculating accuracy
knn_pred = knn_best_model.predict(X_test)
nb_pred = nb_best_model.predict(X_test)
svm_pred = svm_best_model.predict(X_test)

# Displays accuracy results and classification reports
print("KNN Accuracy:", accuracy_score(y_test, knn_pred))

print("\nNaive Bayes Accuracy:", accuracy_score(y_test, nb_pred))

print("\nSVM Accuracy:", accuracy_score(y_test, svm_pred))
```

Figure 3. 10 Implementation Model to Machine Learning

## 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.

# CHAPTER 4

# INITIAL FINDING AND RESULTS

## 4.1 Introduction

This chapter discusses the results and sentiment analysis of the free meal program. This chapter begins with the identification of the dataset, and continues with the results of calculating the proportion of data, creating models and implementing models using machine learning techniques. The machine learning techniques used are K-nearest neighbors (KNN), Naive Bayes and Support Vector Machine (SVM). Based on the results of the implementation of these machine learning techniques, it was found that the KKN and Naive Bayes techniques had a higher percentage of accuracy and classification results compared to SVM. Details of the results and analysis are presented in the following subsections.

## 4.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is very important to do before the modeling stage. Exploratory Data Analysis (EDA) can be briefly interpreted as a process of understanding data to obtain as much information as possible. In addition, EDA can also be done to understand data patterns. The `full_text` column describes the public's reaction on social media X to the free meal program. Then the reaction will be analyzed to obtain the results of sentiment analysis of the program whether it is positive, negative or neutral.

### 4.2.1 Data Collection

The data collection process was carried out using the crawling method using the Python programming language to retrieve data from the social media application

X (formerly known as Twitter). The data that was successfully obtained included various tweets related to the free meal program which was the focus of the study. After the data was collected, the data was saved in CSV file format to facilitate further analysis. The data that had been stored then went through the pre-processing and data cleaning stages, where steps such as removing URLs, user tags (@username), hashtags, numbers, punctuation, and extra spaces were carried out. In addition, stemming was also carried out to return words to their basic form and translation of English words into Indonesian to ensure the data was clean, consistent, and ready for further sentiment analysis. Figure 3.1 illustrates the flowchart process for crawling data.

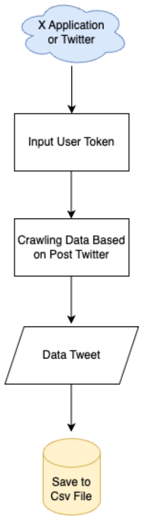
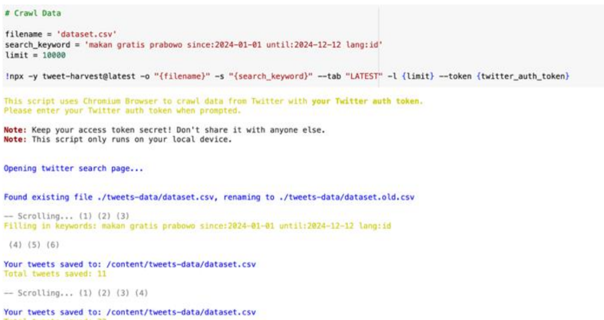



Figure 4. 1 Flowchart of Crawling Data Process

In Figure 4.2 is the data crawling process using Python syntax and utilizing user tweet posts.



```
# Crawl Data

filename = 'dataset.csv'
search_keyword = 'makan gratis prabowo since:2024-01-01 until:2024-12-12 lang:id'
limit = 10000

!npx -y tweet-harvest@latest -o "{filename}" -s "{search_keyword}" --tab "LATEST" -l {limit} --token {twitter_auth_token}

This script uses Chromium Browser to crawl data from Twitter with your Twitter auth token.
Please enter your Twitter auth token when prompted.

Note: Keep your access token secret! Don't share it with anyone else.
Note: This script only runs on your local device.

Opening twitter search page...

Found existing file ./tweets-data/dataset.csv, renaming to ./tweets-data/dataset.old.csv

-- Scrolling... (1) (2) (3)
Filling in keywords: makan gratis prabowo since:2024-01-01 until:2024-12-12 lang:id

 (4) (5) (6)

Your tweets saved to: /content/tweets-data/dataset.csv
Total tweets saved: 11

-- Scrolling... (1) (2) (3) (4)

Your tweets saved to: /content/tweets-data/dataset.csv
```

Figure 4. 2 Process Crawling For Collection Data

In this figure 4.3 is after the data crawling process is successful, then the data will be saved in a file with csv format. The library used for this process is the Pandas library in python. After that the data that has been saved will be displayed in the form of a data frame.

```python
import pandas as pd

# Specify the path to your CSV file
file_path = f"tweets-data/{filename}"

# Read the CSV file into a pandas DataFrame
df = pd.read_csv(file_path, delimiter=",")

# Display the DataFrame
display(df)
```

| | conversation_id_str | created_at | favorite_count | full_text | id_str | image_url | in_reply_to_screen_name | lang | location | quote_count | re |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1866872716541972550 | Wed Dec 11 23:25:45 +0000 2024 | 4 | @pramudyawdynto Target makan gratis e prabowo | 1866987767919677789 | NaN | pramudyawdynto | in | Vladivostok, Russia | 0 | |
| 1 | 1866464297657975123 | Wed Dec 11 21:52:21 +0000 2024 | 1 | @Yudhi2024 @prabowo masih enak dan bergizi mak... | 1866964263254032437 | NaN | Yudhi2024 | in | NaN | 0 | |
| 2 | 1866464297657975123 | Wed Dec 11 21:21:34 +0000 2024 | 0 | @StevanFirman15 @prabowo Yakin pak makan grati... | 1866956515414094156 | NaN | StevanFirman15 | in | Bekasi Barat, Indonesia | 0 | |
| 3 | 1866754957552259364 | Wed Dec 11 18:42:05 +0000 2024 | 0 | @03__nakula Makanya program Prabowo makan sian... | 1866916381872230896 | NaN | 03__nakula | in | Bandung, Jawa Barat | 0 | |
| 4 | 1866877557489926164 | Wed Dec 11 16:07:49 +0000 2024 | 0 | Prancis Dukung Program Makan Bergizi Gratis Pr... | 1866877557489926164 | NaN | NaN | in | Jakarta, Indonesia | 0 | |

Figure 4. 3 Saving and Displaying Data After Data Collection Process

The total data collected is 2916 and 15 columns. The data will be subjected to data pre-processing and data cleaning, so that it will get better and more accurate data.

### 4.2.2 Data Preparation and Cleaning

Data cleaning is an important process in sentiment analysis, especially to ensure that the data used is clean, relevant, and can be processed well by the model. Here are the data cleaning steps carried out on the Twitter tweet dataset about the Prabowo-Gibran free meal program : Initialize Sastrawi Stemmer, Convert text to Lowercase, RemoveURLs, Remove @username, Remove #Hastags, remove numbers, remove punctuation, Extra whitespace, Apply Stemming and Translate data to minimize English word. In figure 4.4 is the python syntax used for data pre-processing.

```python
# 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 4. 4 Syntax for Pre-processing Data

### 4.2.3 Demographic and Distribution Data

Demography is the study of human populations and changes in their quantity as they relate to migration, fertility, and mortality. The term demography is derived from the Greek word and means "describing people." Thus, this discipline deals with the characteristics of populations by considering features such as sex ratio, age structure, composition, spatial distribution, and population density. (Klimczuk, 2021)

In this project, demographic data consists of user data and also user tweet posts related to the free meal program from President Prabowo - Gibran. The following is an explanation of the data distribution in the free food program dataset as shown in Figure 4.5.

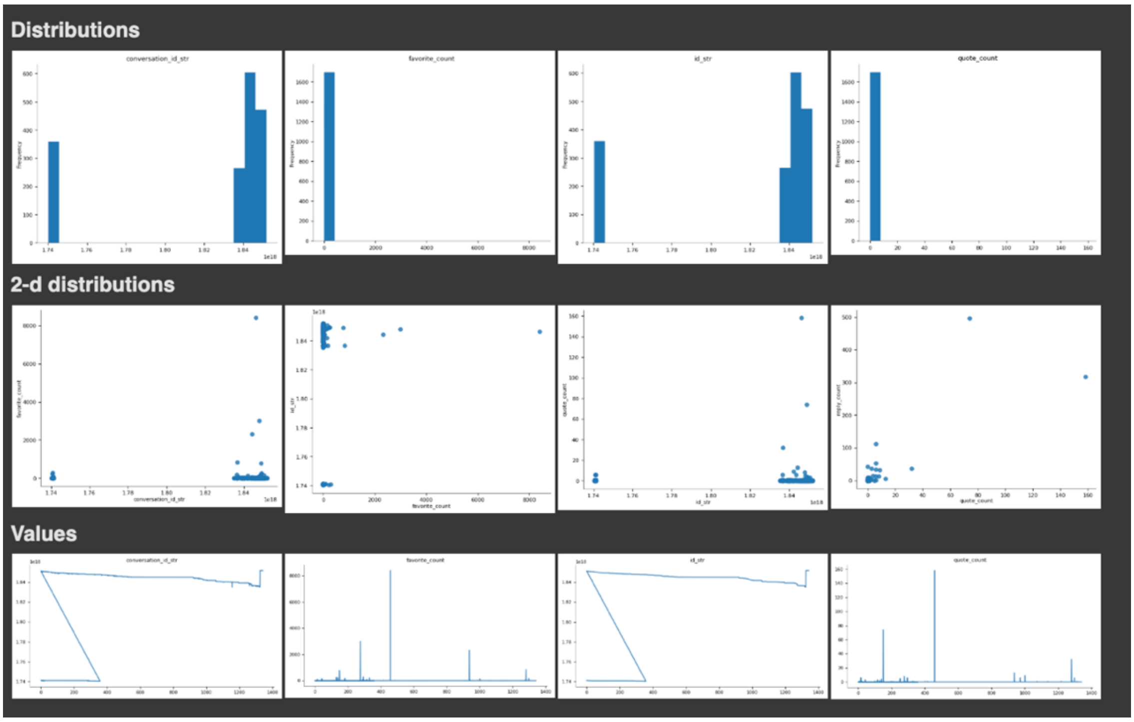



Figure 4. 5 Distribution of data in each column in the dataset

Table 4.1 explains the distribution of data related to each item value in the dataset.

Table 4. 1 Analysis of each column in the dataset

| | |
|---|---|
| **Distributions**<br>This section shows the distribution of values for each column of the dataset. | |
| `conversation_id_str` | This distribution shows unique conversation IDs that are mostly distributed in a certain range. Large ID values |

| | |
|---|---|
| | indicate that this is data taken from Twitter, as IDs are usually long numbers. |
| `favorite_count` | Distribution of the number of "likes" or "favorites" on tweets. Most tweets have a low "like" value (close to zero), indicating that many tweets receive little attention or interaction. |
| `id_str` | Like `conversation_id_str`, this is a unique ID for a tweet. Its distribution follows a similar long ID pattern. |
| `quote_count` | Distribution of the number of "quote retweets". Most of the data has a value of zero, indicating that most tweets are not quoted by other users. However, there are some extreme values with higher "quote" numbers. |
| **2-d Distributions**<br>This section shows the relationship between variables with a 2-dimensional distribution. | |
| `favorite_count vs conversation_id_str` | This graph shows that the number of “likes” is sporadically distributed across the conversation IDs. Most of the “like” values are low, with a few outliers having high “like” counts. |
| `favorite_count vs id_str` | Similar to the previous relationship, but focused on the unique ID of each tweet. The pattern is similar, with a few dots indicating popular tweets. |
| `quote_count vs id_str` | Most tweets have a low quote value, but there are a few outliers where tweets have a significant number of quotes. This suggests that only a small number of tweets attract the attention of other users to re-comment. |
| **Values**<br>This section visualizes the distribution of values in the form of a line: | |
| `conversation_id_str` | The lines indicate sequential IDs. This confirms that the data may have been collected chronologically. |
| `favorite_count` | The distribution pattern shows that most values are close to zero with a few peaks (outliers). |
| `quote_count` | Most of the values are close to zero, indicating tweets that are rarely requoted, but there are a few peaks with higher values. |

In Figure 4.6 is each column in the dataset and also the data type used. It can be seen that all columns are non-null, consisting of 8 objects and 7 int64.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2916 entries, 0 to 2915
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   conversation_id_str     2916 non-null   int64 
 1   created_at              2916 non-null   object
 2   favorite_count          2916 non-null   int64 
 3   full_text               2916 non-null   object
 4   id_str                  2916 non-null   int64 
 5   image_url               692 non-null    object
 6   in_reply_to_screen_name 1002 non-null   object
 7   lang                    2916 non-null   object
 8   location                1352 non-null   object
 9   quote_count             2916 non-null   int64 
 10  reply_count             2916 non-null   int64 
 11  retweet_count           2916 non-null   int64 
 12  tweet_url               2916 non-null   object
 13  user_id_str             2916 non-null   int64 
 14  username                2916 non-null   object
dtypes: int64(7), object(8)
memory usage: 341.8+ KB
```

```
data.columns
```

```
Index(['conversation_id_str', 'created_at', 'favorite_count', 'full_text',
       'id_str', 'image_url', 'in_reply_to_screen_name', 'lang', 'location',
       'quote_count', 'reply_count', 'retweet_count', 'tweet_url',
       'user_id_str', 'username'],
      dtype='object')
```

Figure 4. 6 Dataset Information

In Figure 4.7 Dataset Description, there are extreme values (outliers) in `favorite_count`, `quote_count`, `reply_count`, and `retweet_count` indicating that some tweets are very viral, which may be caused by content factors or accounts with many followers.

```
data.describe()
```

| | conversation_id_str | favorite_count | id_str | quote_count | reply_count | retweet_count | user_id_str |
|---|---|---|---|---|---|---|---|
| **count** | 2.916000e+03 | 2916.000000 | 2.916000e+03 | 2916.000000 | 2916.000000 | 2916.000000 | 2.916000e+03 |
| **mean** | 1.839460e+18 | 13.176269 | 1.839570e+18 | 1.172154 | 2.895405 | 3.250000 | 1.219714e+18 |
| **std** | 3.919510e+16 | 210.665196 | 3.919693e+16 | 30.272856 | 43.586168 | 61.184726 | 6.873334e+17 |
| **min** | 1.740213e+18 | 0.000000 | 1.740399e+18 | 0.000000 | 0.000000 | 0.000000 | 1.538445e+07 |
| **25%** | 1.843504e+18 | 0.000000 | 1.843553e+18 | 0.000000 | 0.000000 | 0.000000 | 8.646796e+17 |
| **50%** | 1.848229e+18 | 0.000000 | 1.848231e+18 | 0.000000 | 0.000000 | 0.000000 | 1.598136e+18 |
| **75%** | 1.862699e+18 | 0.000000 | 1.862782e+18 | 0.000000 | 0.000000 | 0.000000 | 1.696032e+18 |
| **max** | 1.877502e+18 | 8417.000000 | 1.877504e+18 | 1353.000000 | 1708.000000 | 2966.000000 | 1.865041e+18 |

Figure 4. 7 Dataset Description

### 4.2.4 Data Proportion

Proportion data is used to help understand the balance of data between relevant categories (Normal) and categories that can be considered noise (Bot). Duplicated tweets (Bot) are often less relevant for sentiment analysis because they tend not to represent the user's original opinion.

Some of the reasons why proportion data is needed include for the process of Data Quality Identification, Data Cleaning, Sentiment Model Evaluation and Better Decision Making. In the Figure 4.8 it can be seen that the data categories are divided into two, namely Bot and Normal. The distribution of the proportion of each of these data is Bot, there are 14.09% of the total dataset or around 411 data, while the Normal category is the majority of the data, which is around 85.91% or 2505 data. In this case, it can be seen that if the proportion of Bot is too large, the sentiment model can be biased because duplicate tweets have a repetitive pattern and do not reflect real opinions. In this case, the proportion of Bot is 14.09%, which is still considered reasonable to overcome.

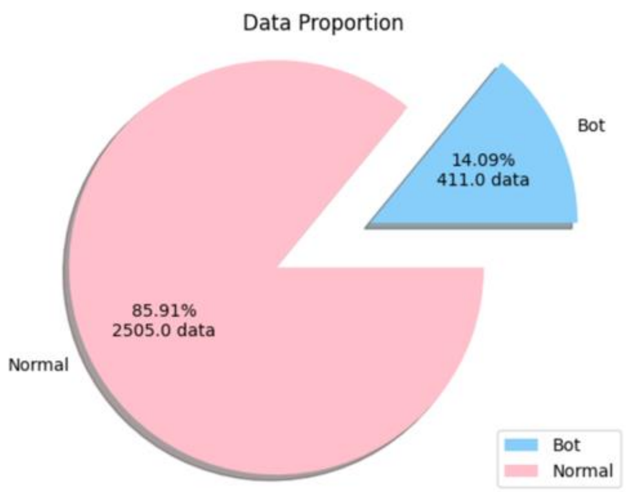



Figure 4. 8 Data Proportion

From the proportion data results, it was found that:

a. Balanced Distribution

   A much larger proportion of the Normal category (82.66%) indicates that the majority of data comes from unique sources, so this dataset is valid enough to be used in sentiment analysis.

b. Bot Influence

(a) Although the number of Bots is small, this data still needs to be considered because it can affect the final results of the analysis if not filtered properly.

c. Initial Conclusion

This graph shows that the dataset is of sufficient quality because most of the data are unique tweets (Normal), but preprocessing steps to eliminate duplication (Bot) are still needed before further analysis.

## 4.3 Sentiment Analysis

In this sentiment analysis section, we will identify each word from social media tweets and categorize whether the word is positive, negative or neutral. Here are some examples of positive, negative and neutral sentences.

Table 4. 2 Some Examples of Sentiment Analysis Sentences

| Full_text | Sentiment |
|---|---|
| sindir keras mahfud md soal program makan sian.. | Positif |
| saltingan banget dia kenapa sihhh dari dulu ga.. | Neutral |
| banyak yang belum sadar bahwa presiden saat in... | Negative |

The Figure 4.9 shows a word cloud for positive sentiment reviews. The word cloud analysis illustrates that “healthy”, “economic movement”, “prosperous”, “thank you”, “future”, “lunch”, “free lunch”, “industrial sector” and “great potential” are the most frequently used words in the reviews. The word “Lunch” is the most frequently used word, indicating that this program often receives positive reviews from public reactions on social media. While the words “healthy” and “prosperous” are the words that are the hopes of this free meal program when it is run by the government under the leadership of President and Vice President Prabowo Subianto and Gibran.


Figure 4. 9 World Cloud of Positive Sentiment

The Figure 4.10 shows a word cloud for negative sentiment reviews. The word cloud analysis illustrates that "no", "prabowo", "problem", "fail", "stupid", "sarcasm", "poor", "corruptor" and "politics" are the most frequently used words in the reviews. The words "fail", "problem" and "poor" illustrate that many people will be pessimistic about this free meal program. They assume that this program will fail and not continue and will not be on target. So it will add new problems for the country, while currently those who need to be helped are people with poor economies to be more prosperous.


Figure 4. 10 World Cloud of Negative Sentiment

The Figure 4.11 shows a word cloud for neutral sentiment reviews. The word cloud analysis illustrates that “nutrition”, “confused”, “program”, “campaign promise”, “help”, “realization”, “children”, “economy” and “need” are the most frequently used words in the reviews. These words may indicate that the meal program provides a little hope for children to get better nutrition. In addition, the community also hopes that the government can realize the program, not just make promises during the campaign.

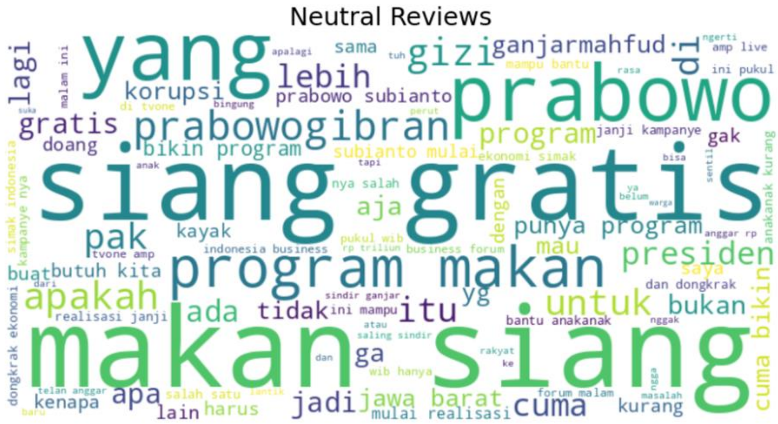



Figure 4. 11 World Cloud of Neutral Sentiment

From the figure 4.12 and 4.13 the results of the sentiment analysis above, it can be seen that the sentiment results are positive. With a total of 95.57% data or 1624 data. Then the neutral data is 2.82% and the least is data with negative sentiment which is only 1.7%. Thus, we can conclude that the results of the sentiment analysis related to the free meal program are positive and can be interpreted as meaning that the public agrees to the free meal program.

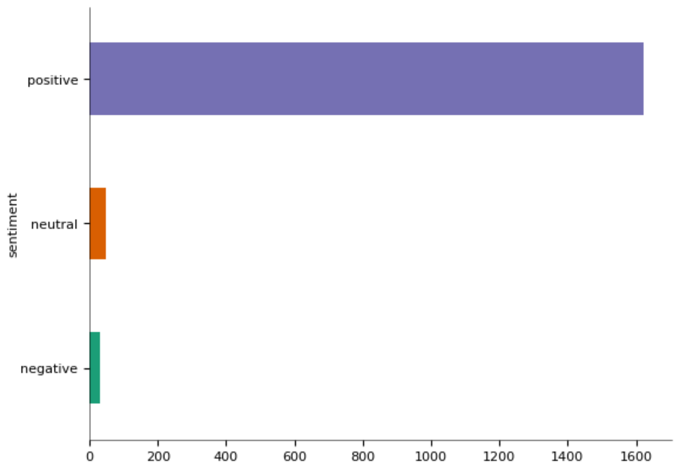



Figure 4. 12 Distribution of Sentiment Analysis Result by Categories

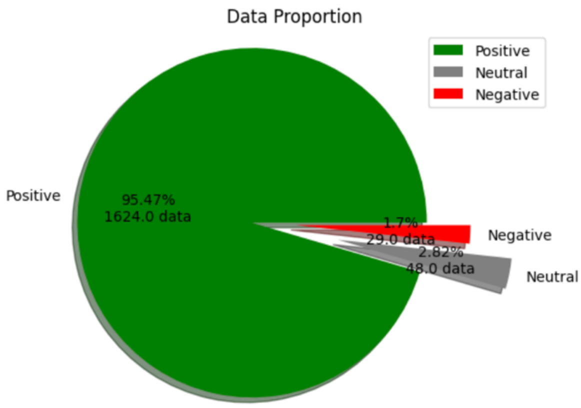



Figure 4. 13 Graph of Proportion Data From Sentiment Analysis Results

## 4.4 Feature Extraction

The feature extraction process is carried out to convert raw text data into numeric representations that can be processed by the machine learning model. The first step is to perform class balancing by ensuring that the amount of data for each sentiment class (positive, neutral, and negative) has the same proportion. This aims to avoid model bias towards classes with dominant data amounts. In this implementation, the under sampling method is applied, namely selecting the same number of samples based on the smallest number of classes, resulting in a balanced distribution with each class having 91 samples.

The next step, the text data is converted into a vector using the TF-IDF (Term Frequency-Inverse Document Frequency) Vectorizer, with a maximum feature limit used of 5,000 features. TF-IDF calculates the weight value of each word based on its frequency of occurrence in a particular document (term frequency) and how unique the word is compared to other documents (inverse document frequency). This process produces a matrix with a size of (273, 1736), where 273 is the number of documents (data) after balancing, and 1736 is the number of unique features generated.

The last step is to encode the sentiment label using `LabelEncoder`. The sentiment labels “positive,” “neutral,” and “negative” are transformed into their respective numeric values (0 for negative, 1 for neutral, and 2 for positive). This process ensures that the target data conforms to a format acceptable to the machine learning algorithm. This transformation produces data ready for the model training process, with numeric representations of the text and encoded sentiment labels.

### 4.5 Model Development

At this stage, we will help the model become Model X and Y. And then use the TF-IDF Vectorizer (Term Frequency-Inverse Document Frequency Vectorizer) technique to convert text data into numeric representations (vectors) before the data is processed using machine learning algorithms. Its main function is to give weight to each word in the document, so that relevant words are more prominent and common words are less influential. TF-IDF Vectorizer is an important step in the text-based machine learning pipeline. It helps capture important information from text data and ignores irrelevant information, thereby improving the performance and accuracy of the prediction model.

The Importance of TF-IDF in Machine Learning :

a) Reducing Data Dimensionality: TF-IDF allows the use of only relevant words (for example, with `max_features=5000`) without having to process all the words in the dataset, making the model more efficient.

b) Reducing Overfitting: Very common or irrelevant words (stopwords) are given low weight or ignored, which helps the model not to be affected by noise.

c) Highlight Relevant Words: Words that are specific and relevant to a particular class will get higher weights, increasing the model's accuracy in understanding the relationship between words and target labels.

d) Compatible with Machine Learning: Machine learning models like SVM, Naive Bayes, or KNN can only process numeric data. TF-IDF converts text data into a numeric format that is acceptable to the model.

In Figure 4.14 is the code syntax for the model creation process and also the implementation of the TF-IDF Vectorizer technique for each 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

# Divide the data into features (X) and targets (y)
X = data['full_text']
y = data['sentiment']
```

```
# TF-IDF Vectorizer to convert text to vectors
vectorizer = TfidfVectorizer(max_features=5000)  # Adjust max_features as needed
X_vectorized = vectorizer.fit_transform(X)
```

```
# Splitting data into training and test data
X_train, X_test, y_train, y_test = train_test_split(X_vectorized, y, test_size=0.2, random_state=42)
```

Figure 4. 14 Creating Data Model and Implement TF-IDF

## 4.6 Model Evaluation and Implement Machine Learning Technique

The model that has been successfully created previously will then be processed for implementation in machine learning techniques. To get better accuracy, in this project researchers will use several machine learning techniques such as KNN (K-Nearest Neighbors), Naive Bayes, and SVM (Support Vector Machine). In this process, Hyperparameter Tuning will be used for each machine learning technique. Hyperparameter tuning is done to find the best parameters of the three machine learning models. Here are the details:

1) KNN (K-Nearest Neighbors)

   Tested parameters:

   a) `n_neighbors`: Number of neighbors considered (3, 5, 7, 9).

   b) weights: How to give weight to neighbors (uniform for all neighbors have the same weight, and distance for weights based on distance).

      The best model is stored in the variable `knn_best_model`.

2) Naive Bayes

   Tested parameters:

   alpha: Smoothing parameter (tested values: 0.1, 0.5, 1.0, 1.5, 2.0).

The best model is stored in the variable `nb_best_model`.

3) SVM (Support Vector Machine)

Tested parameters:

C: Regularization parameter (tested values: 0.1, 1, 10, 100).

kernel: Kernel function (linear for linear kernel, and rbf for radial basis function kernel).

The best model is stored in the variable `svm_best_model`.

After finding the best model for each algorithm, an evaluation is carried out using test data (`X_test` and `y_test`) with the following steps:

a) Test Data Prediction

KNN: Using `knn_best_model` to predict.

Naive Bayes: Using `nb_best_model` to predict.

SVM: Using `svm_best_model` to predict.

b) Calculating Accuracy

Accuracy is calculated with the accuracy score function, which compares the predictions to the original labels in the test data.

After conducting model evaluation and data test, the results were obtained as:

KNN : 94.34%

Naive Bayes : 94.00%

SVM : 94.34%

```
# Predicting test data and calculating accuracy
knn_pred = knn_best_model.predict(X_test)
nb_pred = nb_best_model.predict(X_test)
svm_pred = svm_best_model.predict(X_test)

# Displays accuracy results and classification reports
print("KNN Accuracy:", accuracy_score(y_test, knn_pred))

print("\nNaive Bayes Accuracy:", accuracy_score(y_test, nb_pred))

print("\nSVM Accuracy:", accuracy_score(y_test, svm_pred))
```

```
KNN Accuracy: 0.9434931506849316

Naive Bayes Accuracy: 0.940068493150685

SVM Accuracy: 0.9434931506849316
```

Figure 4. 15 Accuracy Results and Classification Reports

## 4.7 Summary

After conducting several processes for sentiment analysis, model evaluation, and data testing with machine learning data techniques, it was found that the sentiment obtained was positive with a total of 95.47% or 1624 data. And for the accuracy of the machine learning technique, KNN and SVM have the same high accuracy (94.35%) and are better than Naïve Bayes (94.00%) in predicting sentiment. Hyperparameter tuning successfully improves model performance by selecting the best combination of parameters for each algorithm. All models show very good performance with accuracy above 94%, which means that the data has been processed well (for example: through TF-IDF Vectorizer and balanced training-test data division). Although Naïve bayes is slightly lower in accuracy, this algorithm is usually more stable for data with high dimensions and complex distributions, so it can be considered for larger or more varied datasets. With these results, KNN or SVM can be selected as the best model for sentiment analysis cases based on performance on test data.

# CHAPTER 5

# DISCUSSION AND FUTURE WORK

## 5.1 Introduction

This chapter summarizes the results of the sentiment analysis project related to the free meal program carried out by Prabowo-Gibran through crawling data from Twitter or the X application. The results and insights obtained after going through a series of data analysis stages, starting from data collection, cleaning, and applying various machine learning algorithms, provide an overview of public perception of this program. In addition, it also shows an overview of future project development, possibilities for making improvements in terms of quality and accuracy for better analysis. Thus, this study takes a structured approach starting from data processing to model evaluation, with the aim of making a positive contribution to the data-based decision-making process in the Indonesian socio-political order.

## 5.2 Summary

Sentiment Analysis of the Free Meal Program voiced by Prabowo-Gibran tries to get public responses to the program. This analysis clarifies public perception of the policy through data taken from tweets on Twitter through the web scraping process. This project involves several phases, from data collection to final analysis. The data after being scraped will go through a cleaning stage which means that various preprocessing techniques can be carried out, including converting text to lowercase, removing URLs, usernames, punctuation, and irrelevant words. Then, stemming and foreign word translation are carried out to create uniform data.

The cleaned data is then converted into numeric form using the TF-IDF Vectorizer. This technique allows us to represent text in vector form so that it can be used by machine learning algorithms. The three algorithms used in this project are K-

Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machine. The results of the analysis show that KNN and Naive Bayes have the highest level of accuracy, which is 97.61%, while for SVM, the accuracy is slightly lower at 96.80%. This means that the results show that the machine learning model is reliable in analyzing public sentiment based on social media text. With a deeper analysis, it can be seen that the trend of public opinion tends to be positive, which means that the public is still accepted in the 'Free Meals' initiative. It cannot be ignored that there are a small number of people who express negative and neutral sentiments, reflecting concerns or misunderstandings about the program.

From the success of the project, we can draw the following conclusions:

a) Data quality is crucial: Good data cleaning will play a big role in achieving better results after analysis.

b) Appropriate model selection: KNN and Naive Bayes performed very well, so they are the primary choices for this case.

c) Dominantly positive sentiment: The Free Meal Program has received strong positive support from the public.

Overall, the project successfully achieved its goal of measuring public response to a policy program in a structured and data-driven manner. The project also demonstrated that social media sentiment analysis can act as a powerful tool in evaluating public policies in real time.

## 5.3 Future Works

While this project has provided a lot of insights, there are several areas that can be further developed to improve the quality of analysis in future. Some suggestions for future work are as follows:

a) Larger Data Volume

In the current project, the dataset used consists of only a few tweets from Twitter. This data can also be taken from other social media platforms such as Facebook, Instagram, or TikTok for a broader analysis to provide a broader view of public sentiment.

b) Larger Demographic Analysis

This can be extended by integrating demographic information, such as age, location, and gender of users. This will help in providing insights into the variation in public responses across a group of people.

c) Using Deep Learning-Based Model

Traditional models were used in this work, but deep learning-based models using LSTM or Transformer—for example, BERT—will provide better accuracy by capturing complex contextual information from text in future approaches.

d) Clarity in Models

In data-driven decision making, interpretable models must be developed. Further research could be conducted to understand why the model makes certain predictions, making the results of the analysis more understandable to policy makers.

The above steps will allow further research to increase the scope of this project, improving the accuracy and relevance of the results. The current project has paved the way for the use of social media data as an effective public policy evaluation tool; thus, further development will have greater implications in the future for strategic decision making.

## REFERENCES


Adak, A., Pradhan, B., & Shukla, N. (2022). Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review. *Foods*, *11*(10). https://doi.org/10.3390/foods11101500

Ainin, S., Feizollah, A., Anuar, N. B., & Abdullah, N. A. (2020). Sentiment analyses of multilingual tweets on halal tourism. *Tourism Management Perspectives*, *34*. https://doi.org/10.1016/j.tmp.2020.100658

Ajhari, A. A. (2023). The Comparison of Sentiment Analysis of Moon Knight Movie Reviews between Multinomial Naive Bayes and Support Vector Machine. *Applied Information System and Management (AISM)*, *6*(1), 13–20. https://doi.org/10.15408/aism.v6i1.26045

Awazon. (2022). *Apa yang dimaksud dengan Analisis Sentimen?* Https://Aws.Amazon.Com/Id/What-Is/Sentiment-Analysis.

Hudaefi, F. A., Caraka, R. E., & Wahid, H. (2022). Zakat administration in times of COVID-19 pandemic in Indonesia: a knowledge discovery via text mining. *International Journal of Islamic and Middle Eastern Finance and Management*, *15*(2), 271–286. https://doi.org/10.1108/IMEFM-05-2020-0250

Klimczuk, A. (2021). Introductory Chapter: Demographic Analysis. In *Demographic Analysis - Selected Concepts, Tools, and Applications*. IntechOpen. https://doi.org/10.5772/intechopen.100503

Kolchyna, O., Souza, T. T. P., Treleaven, P., & Aste, T. (2015). *Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination*. http://arxiv.org/abs/1507.00955

Landwehr, P. M., Wei, W., Kowalchuck, M., & Carley, K. M. (2016). Using tweets to support disaster planning, warning and response. *Safety Science*, *90*, 33–47. https://doi.org/10.1016/j.ssci.2016.04.012

Monselise, M., Chang, C. H., Ferreira, G., Yang, R., & Yang, C. C. (2021). Topics and sentiments of public concerns regarding COVID-19 vaccines: Social media trend analysis. *Journal of Medical Internet Research*, *23*(10). https://doi.org/10.2196/30765

Nguyen, T. T., Criss, S., Allen, A. M., Glymour, M. M., Phan, L., Trevino, R., Dasari, S., & Nguyen, Q. C. (2019). Pride, love, and twitter rants: Combining machine learning and qualitative techniques to understand what our tweets reveal about race in the us. *International Journal of Environmental Research and Public Health*, *16*(10). https://doi.org/10.3390/ijerph16101766

Program, T., Siang, M., Tundo, G., & Rachmawati, D. N. (2024). Implementasi Algoritma Naive Bayes untuk Analisis Sentimen. In *Jurnal Indonesia : Manajemen Informatika dan Komunikasi (JIMIK)* (Vol. 5, Issue 3). https://journal.stmiki.ac.id

Ratna Patria. (2022). *Analisis Sentimen Media Sosial: Pengertian, Manfaat, Cara*. Https://Www.Domainesia.Com/Berita/Analisis-Sentimen-Media-Sosial/.

Sudo, K., Murasaki, K., Kinebuchi, T., Kimura, S., & Waki, K. (2020). Machine Learning–Based Screening of Healthy Meals From Image Analysis: System Development and Pilot Study. *JMIR Formative Research*, *4*(10), e18507. https://doi.org/10.2196/18507

Tao, Y., Zhang, F., Shi, C., & Chen, Y. (2019). Social media data-based sentiment analysis of tourists' air quality perceptions. *Sustainability (Switzerland)*, *11*(18). https://doi.org/10.3390/su11185070