Applying Graph Theory in Decision Trees for Public Sentiment Analysis on Twitter Tweets about the Israel-Palestine Conflict

Rizqika Mulia Pratama - 13522126[1]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]*13522126@mahasiswa.itb.ac.id*

***Abstract*—This paper introduces a novel approach employing graph theory, specifically in the form of decision trees, to determine the sentiment of tweets related to the Israel-Palestine conflict. The relationship between various features extracted from tweets are modeled as a decision tree structure, allowing for the classification of tweets as either pro-Israel or pro-Palestine. By leveraging this graph-based decision methodology, the paper aims to provide an effective means of sentiment analysis on Twitter data related to the Israel-Palestine conflict, shedding light on the prevailing opinions within the online discourse.**

***Keywords*—Graph Theory, Decision Tree, Israel-Palestine conflict, Sentiment Analysis.**

I. INTRODUCTION

The conflict between Israel and Palestine is one of the global issues that has garnered international attention for several decades. In this digital era, social media serves as the primary channel for individuals to voice their opinions, share information, and engage in discussions related to the Israel-Palestine conflict. With the popularity of platforms like Twitter, billions of tweets are posted every day, reflecting diverse perspectives and responses to global issues, including the Israel-Palestine conflict.

This paper aims to develop a graph theory-based method to conduct a sentiment analysis of public tweets discussing the Israel-Palestine conflict on Twitter. This approach combines graph theory concept with artificial intelligence to build a decision tree capable of predicting whether a tweet is pro-Israel, pro-Palestine, neutral, or remains unclassified.

Using data obtained through scraping Twitter, this paper will elucidate the process of constructing the decision tree, implementing graph theory in sentiment analysis, and presenting findings that offer insight into how public sentiments are divided concerning the conflict.

Through this approach, it is hoped that this paper contributes to understanding the dynamics of public discourse on the Israel-Palestine conflict in social media. Additionally, the research findings may pave the way for further exploration in applying graph theory and artificial intelligence in sentiment analysis related to sensitive issues in the digital era.

II. LITERATURE REVIEW

A. Graph Theory

A graph is a collection of points or nodes connected by edges. A graph G = (V, E) is defined, where V is a non-empty set of nodes, denoted as n $\{v_1, v_2, \dots, v_n\}$, and E is a set of edges connecting pairs of nodes $\{e_1, e_2, \dots, e_n\}$.

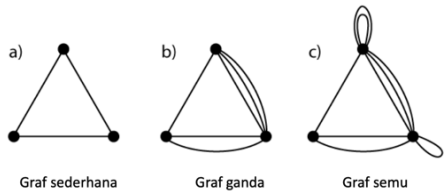


Figure 1. Types of Graphs based on bracelets and double sides.
Source: https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2023-2024/19-Graf-Bagian1-2023.pdf

Based on the presence of loops or multiple edges, a graph can be classified into three types:

1. *Graf sederhana* (simple graph), which does not contain loops or multiple edges.
2. *Graf ganda* (multigraph), which contains multiple edges.
3. *Graf semu* (pseudograph), which contains both loops and multiple edges.

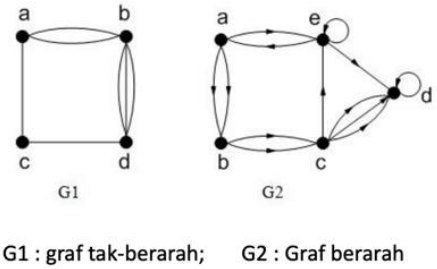


Figure 2. Graph types are based on the direction orientation of the graph.
Source: https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2023-2024/19-Graf-Bagian1-2023.pdf

Based on the orientation of edges, a graph can be categorized into two types:

1. *Graf tak-berarah* (Undirected graph), which does not have a specified direction for its edges.
2. *Graf berarah* (Directed graph), which has a specified direction for its edges.

Here are some terminologies related to graphs:

1. Adjacency, two nodes are considered adjacent if they are connected by an edge.
2. Incident, an edge is said to be an incident to a node if the edge connects the node.
3. Isolated node, an isolated node is a node that has no incident edges.
4. Empty graph, an empty graph is a graph with an empty set of edges.
5. Degree, the degree of a node is the number of edges incident to that node.
6. Path, a path of length n from the initial node v0 to the destination node v_n is a sequence pf alternating node and edges (v_0, e_1, v_1, e_1, … , e_1, v_n) such each $e_1 = (v_0, v_1)$, … , $e_n = (v_{n-1}, v_n)$ is an edge in the graph.
7. Connected, two nodes are considered connected if there is a path from the initial node.

B. Tree

Tree, in graph theory, is a fundamental concept representing connected undirected graphs that are acyclic. A tree is an essential structure in which each node is connected, and there are no cycles within its edges. This property distinguishes trees from other types of graphs and makes them a crucial element in various applications, including computer science, network analysis, and optimization problems.

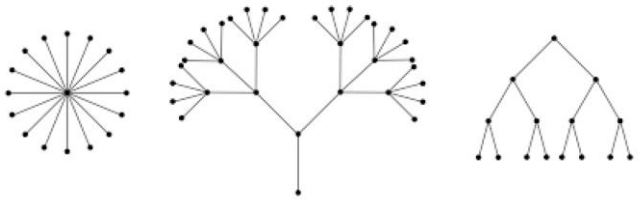


Figure 3. Example of tree.
Source: https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2023-2024/22-Pohon-Bag1-2023.pdf

A tree satisfies three key properties:

1. Connectedness, every node in the tree is reachable from every other node through a path.
2. Acyclicity, the tree contains no cycles, meaning there are no closed paths that start and end at the same node.
3. Undirectedness, edges don't have a specified direction, indicating a symmetric relationship between connected nodes.

C. Decision Tree Algorithm

A decision tree is a supervised learning algorithm that can be used to classify data. It is a tree-like structure that represents a series of decisions that can be made to reach a conclusion.

Decision trees are a popular choice for classification problems because they are relatively easy to understand and interpret. They are also relatively efficient to train and can be used to handle large datasets.

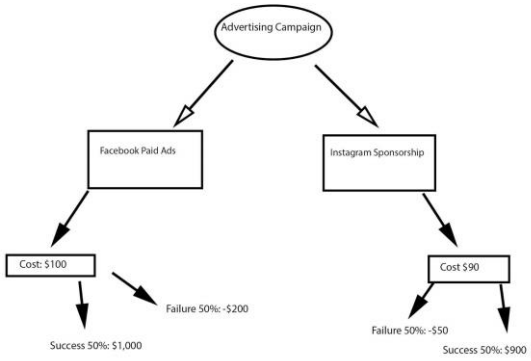


Figure 4. Simple decision tree.
Source: https://venngage.com/blog/what-is-a-decision-tree/

A decision tree is composed of a series of nodes, each of which represents a decision that can be made. The nodes are connected by edges, which represent the possible outcomes of the decisions.

The root node of the tree represents the initial decision that must be made. The child nodes of the root node represent the possible outcomes of that decision. The process continues recursively, with each child node representing a decision that must be made, and so on.

D. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a field of natural language processing (NLP) that focuses on determining the sentiment or emotion expressed in a piece of text. It plays a crucial role in understanding public opinion, customer feedback, and social media interactions.

The process of sentiment analysis involves analyzing text data to categorize it into different sentiment classes, such as positive, negative, or neutral. Machine learning and NLP techniques are commonly employed to automatically classify sentiments based on features like words, phrases, or even contextual information.

III. METHODOLOGY

A. Data Scraping

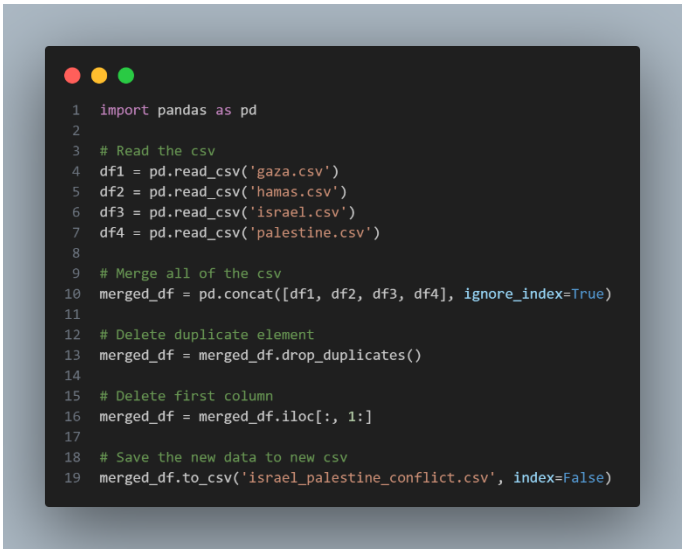
In this phase of the study, data was gathered from Twitter through the utilization of the 'ntscraper' tool, a web scraping tool designed for Twitter data extraction. The primary goal was to compile a dataset consisting of 2000 tweets, with 500 tweets each associated with specific keywords, namely "hamas", "gaza", "palestine", "israel". These keywords were chosen to capture diverse discussions and perspectives related to the Israel-Palestine conflict on Twitter.

Here's the code to scrape the Twitter data:

```python
# Importing necessary libraries
import pandas as pd
from ntscraper import Nitter

# Initializing Nitter scraper
scraper = Nitter()

def get_tweets(query, number_of_tweets, languages):
    # Getting tweets using the Nitter scraper
    tweets = scraper.get_tweets(query, number=number_of_tweets, language=languages)

    # Extracting relevant information from the tweets
    final_tweets = []
    for tweet in tweets['tweets']:
        data = [tweet['user']['username'], tweet['text'], tweet['stats']['likes'], tweet['date']]
        final_tweets.append(data)

    # Creating a DataFrame from the extracted data
    data = pd.DataFrame(final_tweets, columns=['username', 'text', 'likes', 'date'])

    return data

# Getting tweets related to 'hamas' in English and saving to CSV
data = get_tweets('hamas', 500, 'en')
data.to_csv('hamas.csv')
```

Figure 5. Twitter scrapping using Nitter.
Source: personal documentation

The results from all four keyword categories were consolidated into a comprehensive dataset. The Pandas library was employed to efficiently combine the individual datasets into a cohesive CSV file.



```python
import pandas as pd

# Read the csv
df1 = pd.read_csv('gaza.csv')
df2 = pd.read_csv('hamas.csv')
df3 = pd.read_csv('israel.csv')
df4 = pd.read_csv('palestine.csv')

# Merge all of the csv
merged_df = pd.concat([df1, df2, df3, df4], ignore_index=True)

# Delete duplicate element
merged_df = merged_df.drop_duplicates()

# Delete first column
merged_df = merged_df.iloc[:, 1:]

# Save the new data to new csv
merged_df.to_csv('israel_palestine_conflict.csv', index=False)
```

Figure 6. CSV merging using pandas.
Source: personal documentation

In this step, the script combines information from four separate CSV files (‘gaza.csv’, ‘hamas.csv’, ‘israel.csv’, ‘palestine.csv’) into a single file. Duplicate entries are removed to ensure the integrity of the data. The resulting dataset is saved to a new CSV file named ‘israel_palestine_conflict.csv’ with the 2000 total data because there is no duplicate data.

The purpose of this data collection was to create a well-rounded dataset that reflects various perspectives and conversations circulating on Twitter regarding the specified keywords.

Here’s the result in CSV format:

```
israel_palestine_conflict.csv
username,text,likes,date
@MuhammadSmiry,Israel killed 350 Palestinians in the last 24 hours in Gaza  The total without the
@mohammedakunjee,"The cries of the children of Gaza as they died, the screams as their limbs were
@mynameisjro,The killing of Refaat Alareer is a testament that Israel is systematically murdering
@ashoswai,"This is Gaza! If these images have not troubled you for the last two months, this is no
@jacksonhinklle,💔PS Childhood in Gaza under the ISRAELI GENOCIDE!,28138,"Dec 7, 2023 · 5:58 PM UT
@MuhammadSmiry,I am still alive | Gaza,4184,"Dec 8, 2023 · 3:27 AM UTC"
@zaramagnusson,"Only in Gaza do you become an orphan and a parent at the age of 7 💔   These child
@MayaSchenwar,"This Hanukkah, we need to rededicate ourselves to action for a ceasefire and a free
@AkbarSAhmed,"NEW: Gaza may soon become a bigger war - US officials fear Israel wants weapons to f
@Shirinkhan0,No difference  #Gaza_Geniocide,760,"Dec 7, 2023 · 6:24 PM UTC"
@Timesofgaza,Schools Are Being Targeted :  Israeli snipers positioned on top of a school where dis
@OliLondonTV,At least 100 Hamas terrorists captured by IDF Forces today.   Pro-Palestinian account
@visegrad24,Another good day for the IDF  Hamas will soon have lost control of all parts of Gaza,1
@KatanaSpeaks,"TikTok banned @WomenPhilosophers after I posted videos on:  How children are being
```

Figure 7. Twitter scrapping result in CSV format.
Source: personal documentation

B. Data Processing

In the data processing stage, a set of specific keywords is selected to form the basis for constructing the decision tree. These keywords are crucial as they determine criteria for classifying tweets into categories such as “Pro-Israel”, “Pro-Palestine”, or “Unclassified. After selecting the keywords, the decision tree is created using a straightforward process. The script employs a decision tree structure based on a series of conditions that evaluate the presence or absence of specific words in the tweet text. This decision tree serves as the foundation for classifying tweets according to their sentiments related to the Israel-Palestine conflict.

The keywords chosen for constructing the decision tree reflect crucial terms associated with the Israel-Palestine conflict. These keywords are strategically selected to capture different aspects of the discourse. The selected keywords include:

Table 1. *Sentiment analysis categories for Israel-Palestine conflict keywords.*

Keyword	Classification
Free	Unclassified
Palestine	Pro-Palestine
Palestinian	Pro-Palestine
Israel	Pro-Israel
Israeli	Pro-Israel
Hostage	Pro-Israel
Hamas	Pro-Israel
Zionist	Pro-Palestine
Kill	Pro-Palestine
Torture	Pro-Palestine
Humiliate	Pro-Palestine
Murder	Pro-Palestine
Assault	Pro-Palestine
Force	Pro-Palestine
God	Unclassified
Child	Pro-Palestine
Right	Pro-Palestine
Genocide	Pro-Palestine
Justice	Pro-Palestine
Equality	Pro-Palestine
Freedom	Pro-Palestine
Democracy	Pro-Palestine
Oppression	Pro-Palestine
Terrorist	Pro-Israel
Martyred	Pro-Palestine

The next step involves utilizing these keywords to build a decision tree for sentiment analysis. The decision tree is designed to classify tweets related to the Israel-Palestine conflict into categories such as Pro-Israel, Pro-Palestine, or Unclassified. The decision-making process is guided by the presence or absence of specific keywords, reflecting the diverse sentiments and opinions expressed on Twitter.

Here's the illustration of the decision tree in this study:

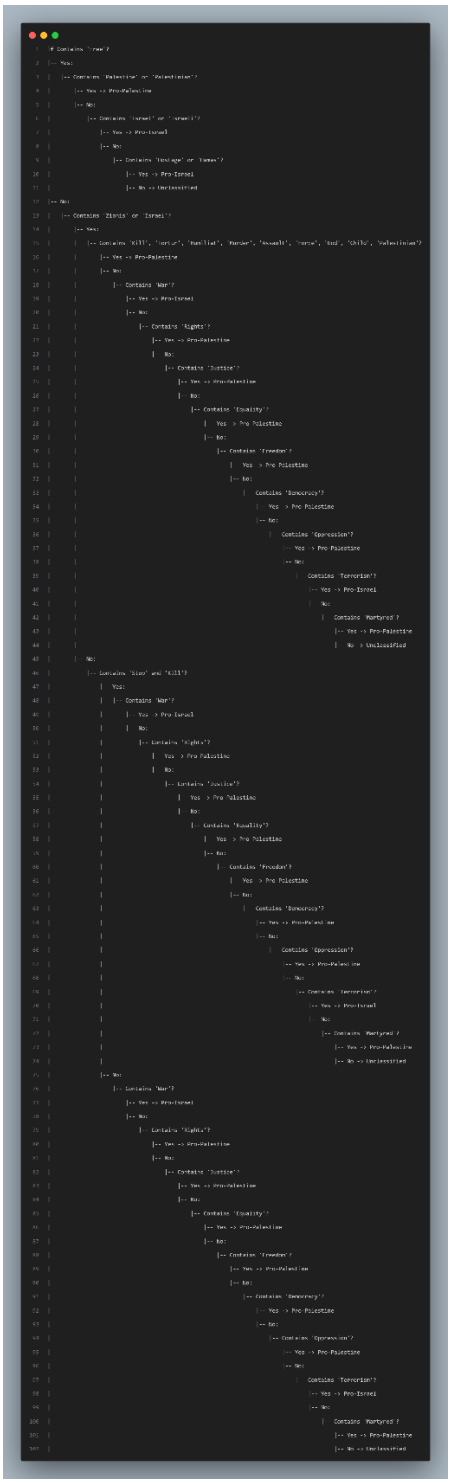


***Figure 8**. Israel-Palestine conflict graph-based decision tree.*
Source: personal documentation

The decision tree above illustrates the classification process based on key keywords. This tree comprises a total of 94 nodes, including both decision nodes and leaves. At the end of the branches, there are 47 leaves, representing the final classification outcomes. Each node corresponds to a decision point based on the presence or absence of specific keywords. The tree is designed to categorize tweets into different classes, such as Pro-Israel, Pro-Palestine, or Unclassified, depending on the content of the tweets.

Additionally, the decision tree is interconnected by a total of 127 edges, guiding the flow of decisions within the structure. These edges play a crucial role in connecting various nodes and leaves, forming a comprehensive network that captures the nuanced classification criteria of the tweets.

The decision tree employs a hierarchical structure, with each level representing a different set of keywords and their associated classifications. The classification outcome is influenced by the nuanced interpretation of the chosen keywords and their combinations. This approach recognizes that the sentiment expressed in tweets can vary based on the context and the interplay of multiple keywords.

After constructing the decision tree, the next step involves classifying tweets using several functions. Each function plays a crucial role in the classification process:

```python
class TreeNode:
    def __init__(self, keyword = None, keyword_list = None, yes_branch = None, no_branch = None, classification = None):
        self.keyword = keyword
        self.keyword_list = keyword_list
        self.yes_branch = yes_branch
        self.no_branch = no_branch
        self.classification = classification

    def check_keyword(self, sentence):
        sentence_lower = sentence.lower()
        if self.keyword:
            return self.keyword.lower() in sentence_lower
        if self.keyword_list:
            return any(kw.lower() in sentence_lower for kw in self.keyword_list)
        return False

def classify(sentence, node):
    if node.classification:
        return node.classification
    if node.check_keyword(sentence):
        return classify(sentence, node.yes_branch)
    else:
        return classify(sentence, node.no_branch)
```

***Figure 9**. TreeNode class implementation for sentiment analysis.*
Source: personal documentation

1. TreeNode Class
   Purpose: Applies the decision tree to classify tweets.
   Description: Each tweet from the 'israel_palestine_conflict.csv' file undergoes the classification process. The classify function is called for each tweet, and the resulting classification label is added as a new column named 'Classification' in the file.
2. Classify Function
   Purpose: Assigns a classification label to a tweet based on the decision tree.
   Description: This recursive function takes a tweet and the current decision tree node as input. It navigates through the decision tree, checking the presence of keywords until a classification label is reached.

3. Data Classification Process

```
df = pd.read_csv('israel_palestine_conflict.csv')

pro_israel = []
pro_palestine = []
unclassified = []


for index, row in df.iterrows():
    sentence = row['text']
    classification = classify(sentence, root)

    if classification == "Pro-Israel":
        pro_israel.append([row['username'], sentence, row['likes'], row['date']])
    elif classification == "Pro-Palestine":
        pro_palestine.append([row['username'], sentence, row['likes'], row['date']])
    else:
        unclassified.append([row['username'], sentence, row['likes'], row['date']])
```

Figure 10. *Data classification process using classify function*.
Source: personal documentation

Purpose: Applies the decision tree to classify tweets.
Description: Each tweet from the 'israel_palestine_conflict.csv' file undergoes the classification process. The classify function is called for each tweet, and the resulting classification label is added as a new column named 'Classification' in the file.

4. Creation of Separate Data Frame

```
# Create DataFrames
df_pro_israel = pd.DataFrame(pro_israel, columns=['username', 'text', 'likes', 'date'])
df_pro_palestine = pd.DataFrame(pro_palestine, columns=['username', 'text', 'likes', 'date'])
df_unclassified = pd.DataFrame(unclassified, columns=['username', 'text', 'likes', 'date'])
```

Figure 11. *Data frame separation.*
Source: personal documentation

Purpose: Organizes tweets into three categories: Pro-Israel, Pro-Palestine, and Unclassified.
Description: The tweets are divided into three separate data frames based on their classification labels. Tweets classified as Pro-Israel are stored in the ‘df_pro_israel’ data frame, Pro-Palestine tweets in the ‘df_pro_palestine’ data frame, and Unclassified tweets in the ‘df_unclassified’ data frame. Each data frame includes relevant information such as the username, tweet text, likes, and date.

IV. Analysis and Result

A. Single Data Analysis

```
Enter Tweet: Israel keeps killing children in gaza everyday.
The sentiment classification for the sentence is: Pro-Palestine
```

Figure 12. *Example of Pro-Palestine tweet.*
Source: personal documentation

In the first example, the output is accurate, categorized as 'Pro-Palestine.' This classification is based on the presence of key terms such as 'Israel' and 'child.' This result is consistent with the decision tree construction, where the sentiment leans towards 'Pro-Palestine' when a tweet contains both 'Israel' and 'Child' concurrently. This rule was established during the decision tree analysis phase."

```
Enter Tweet: Hamas is a terrorist, they took Israeli as a hostage!!
The sentiment classification for the sentence is: Pro-Israel
```

Figure 13. *Example of Pro-Israel tweet.*
Source: personal documentation

For the second example, the phrase "Hamas is a terrorist, they took Israeli as a hostage!!" was input. The correct output is indeed "Pro-Israel." This classification is based on the analysis that tweets containing the words "terrorist" and "Hamas" tend to lean towards a pro-Israel sentiment. Thus, in the decision tree, a rule is established: if a tweet contains "Hamas" and is followed or preceded by the word "terrorist," it is classified as a pro-Israel tweet.

```
Enter Tweet: Gaza is like a prison
The sentiment classification for the sentence is: Unclassified
```

Figure 14. *Example of Unclassified tweet.*
Source: personal documentation

For the last example, given the input tweet “Gaza is like a prison” and classified as “Unclassified,” the tweet should lean more towards pro-Palestine sentiment. However, the type of decision tree developed in this experiment may not capture context effectively as it relies solely on graph theory without a natural language processing algorithm. Thus, it only examines the sequence of words, leading to the classification of pro-Palestine that may be sarcastic or unrelated to the Israel-Palestine conflict in general as “Unclassified.”

A. Dataset Analysis

In the dataset analysis, a check was conducted on 2000 tweets originating from four different search keywords. This data was extracted from the israel_palestine_conflict.csv file, which had been previously processed and separated into the “Pro-Palestine,” “Pro-Israel,” and “Unclassified” data frames. Subsequently, the data was plotted onto a bar chart using the Bokeh library.

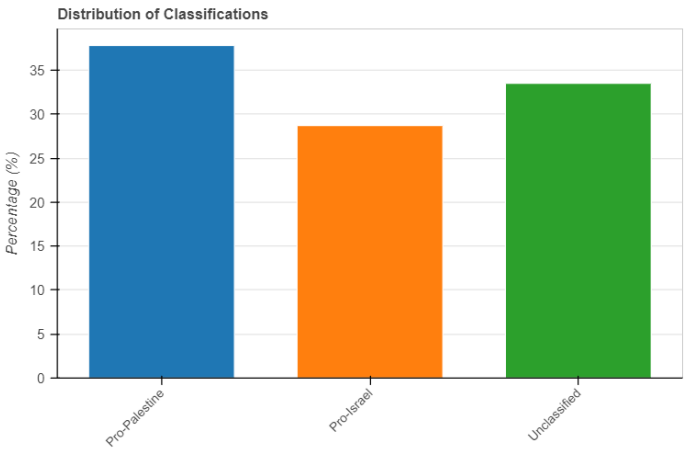


Figure 15. *Comparison of the number of classification results.*
Source: personal documentation

From the processing results using the decision tree created for the entire set of 2000 tweets, it was gathered that 756 tweets tended to be pro-Palestine, 574 tweets tended to be pro-Israel, and 670 tweets were classified as unclassified. The prevalence

of unclassified tweets could be attributed to the decision tree relying solely on graph theory without advanced natural language processing algorithms. Consequently, the decision tree created is not yet capable of understanding the context of a sentence, merely checking for the presence of specific words in the sentence. Tweets classified as "Unclassified" are likely those containing ambiguity, sarcasm, or the use of words irrelevant to the Israel-Palestine conflict.

V. Conclusion and Suggestions

In conclusion, this paper demonstrates the effectiveness of utilizing graph theory to construct decision trees capable of classifying sentiments on Twitter regarding the Israel-Palestine conflict. The chosen keywords, such as "Hamas," "Gaza," "Israel," and "Palestine," revealed a divided sentiment among users, aligning either with the pro-Israel or pro-Palestine stance. Individual biases become apparent through users' tweets on Twitter.

The application of sentiment analysis, particularly in the context of the Israel-Palestine conflict in 2023, holds significant value for various reasons. It provides insights into the dissemination of propaganda on social media and indicates varying levels of support for different sides based on geographical locations. From Figure 15, it is evident that approximately 38% of individuals support Palestine, 29% support Israel, and around 33% remain unclassified. This information allows for an understanding of perspectives from various parts of the world.

The decision tree developed in this experiment can be considered a primitive form of Artificial Intelligence as it can determine sentiment, albeit with some inaccuracies. This is attributed to the experiment's emphasis on graph theory as a discrete mathematics topic without incorporating advanced natural language processing algorithms. Future developments could enhance the decision tree by integrating sophisticated natural language processing algorithms, enabling it to capture the context of sentences, including sarcasm and ambiguity.

VI. Attachment

GitHub link to the source code and data used in the experiment: Sentiment Analysis on Twitter Tweets about the Israel Palestine Conflict

VII. Acknowledgment

The author expresses gratitude to all parties who have assisted in the preparation of this paper, especially to:

1. The Almighty God, for His grace and guidance, allowing the smooth completion of this paper.
2. Both parents for providing both moral and material support to the author.
3. Extended family and friends who have encouraged and aided in the completion of this paper.
4. Dr. Ir. Rinaldi Munir, M.T., and Monterico Adrian, S.T., M.T., as the lecturers for the IF2120 Discrete Mathematics course, for kindly imparting additional knowledge and offering solutions to challenges encountered in writing this paper.
5. Friends in the BYTE cohort, especially those in the IF K3 class, who mutually supported each other during lectures and learned together throughout this semester.

May the Almighty God repay all the kindness that has been bestowed. May this paper serve as a contribution to the academic community.

References

[1] IBM, “What is a Decision Tree | IBM,” *www.ibm.com*. https://www.ibm.com/topics/decision-trees#:~:text=A%20decision%20tree%20is%20a (accessed Dec. 09, 2023).

[2] Amazon, “What is Sentiment Analysis? - Sentiment Analysis Explained - AWS,” *Amazon Web Services, Inc.*, 2023. https://aws.amazon.com/what-is/sentiment-analysis/#:~:text=Sentiment%20analysis%20is%20the%20process (accessed Dec. 09, 2023).

[3] dgnsrekt, “nitter-scraper: Scrape Twitter API without authentication using Nitter.,” *PyPI*. https://pypi.org/project/nitter-scraper/0.4.2/ (accessed Dec. 08, 2023).

[4] “MULTI-LEVEL ANALYSIS OF POLITICAL SENTIMENTS USING TWITTER DATA: A CASE STUDY OF THE PALESTINIAN-ISRAELI CONFLICT - ProQuest,” *www.proquest.com*. https://www.proquest.com/openview/48e4caec43dd93f2a7cb62500ef47919/1?pq-origsite=gscholar&cbl=5500744 (accessed Dec. 07, 2023).

[5] R. Munir, “Graf (Bag.1) Bahan Kuliah IF2120 Matematika Diskrit,” 2023. Accessed: Dec. 08, 2023. [Online]. Available: https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2023-2024/19-Graf-Bagian1-2023.pdf

[6] R. Munir, “Pohon (Bag. 1) Program Studi Teknik Informatika STEI-ITB,” 2023. Accessed: Dec. 08, 2023. [Online]. Available: https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2023-2024/22-Pohon-Bag1-2023.pdf

[7] R. Munir, “Pohon (Bag. 2) Program Studi Teknik Informatika STEI-ITB,” 2023. Accessed: Dec. 08, 2023. [Online]. Available: https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2023-2024/23-Pohon-Bag2-2023.pdf

STATEMENT

I hereby declare that the paper I have written is my own work, not a paraphrase or translation of someone else’s paper, and it is not plagiarized.

Bandung, 10 December 2023

Rizqika Mulia Pratama
13522126