Exploratory Analysis of Clustering Algorithms in Categorizing PH Spam Texts

RHEA SALVE J. GUINGAO
2021 – 2362 | BS COMPUTER SCIENCE IV | CSC172 CS4

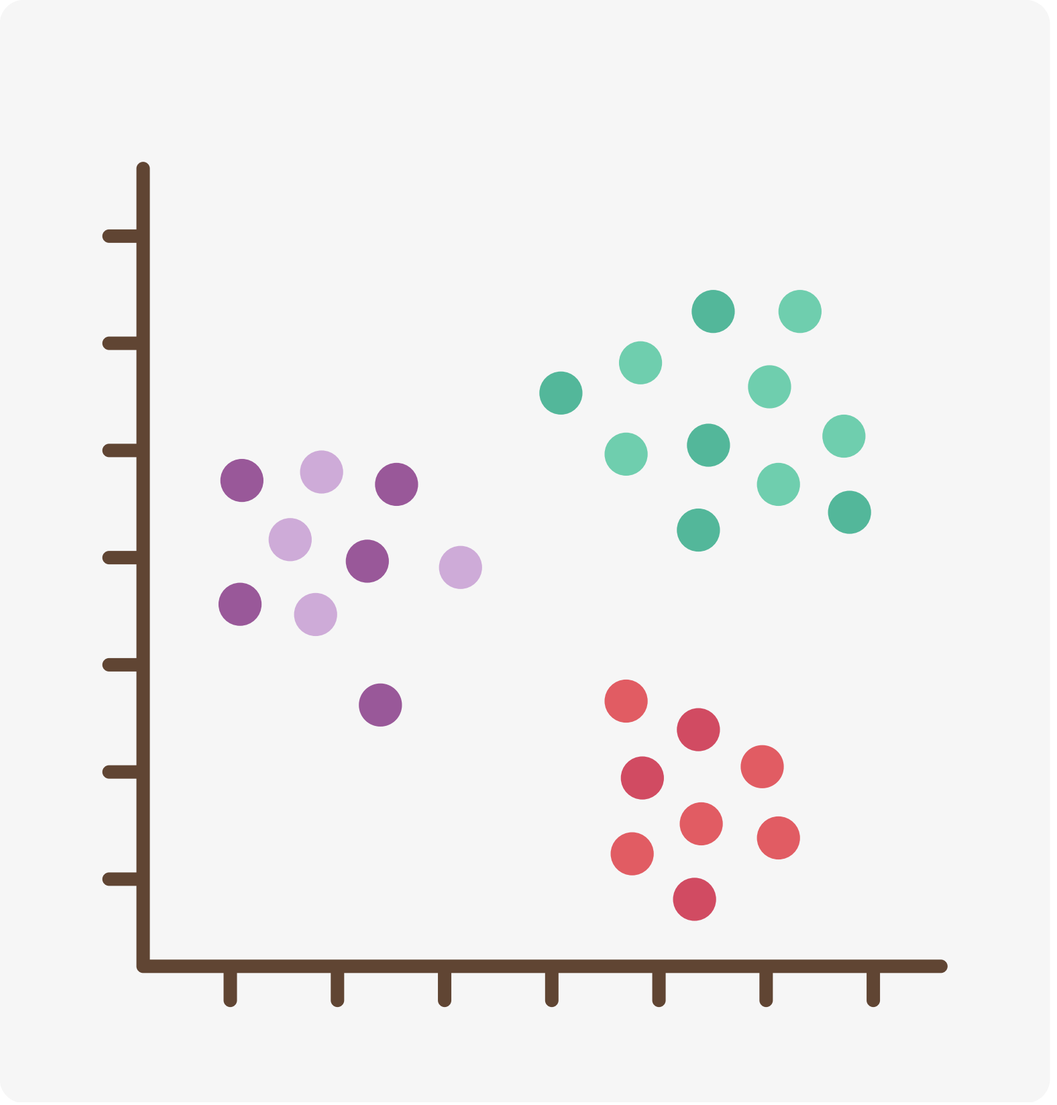
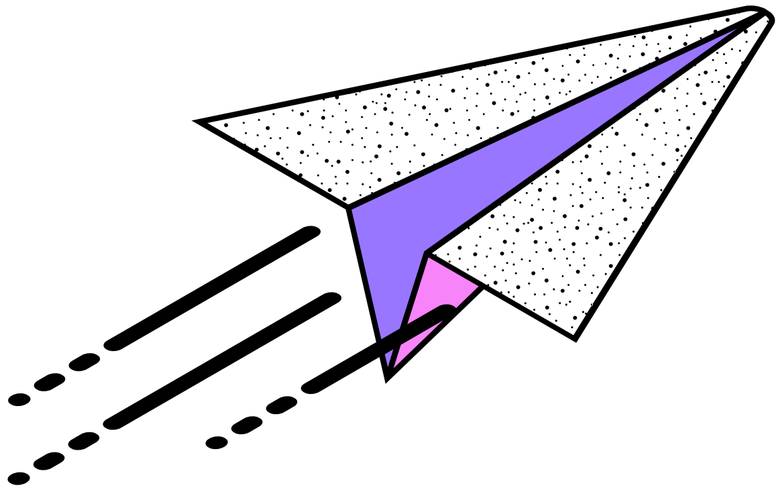


Table of Contents



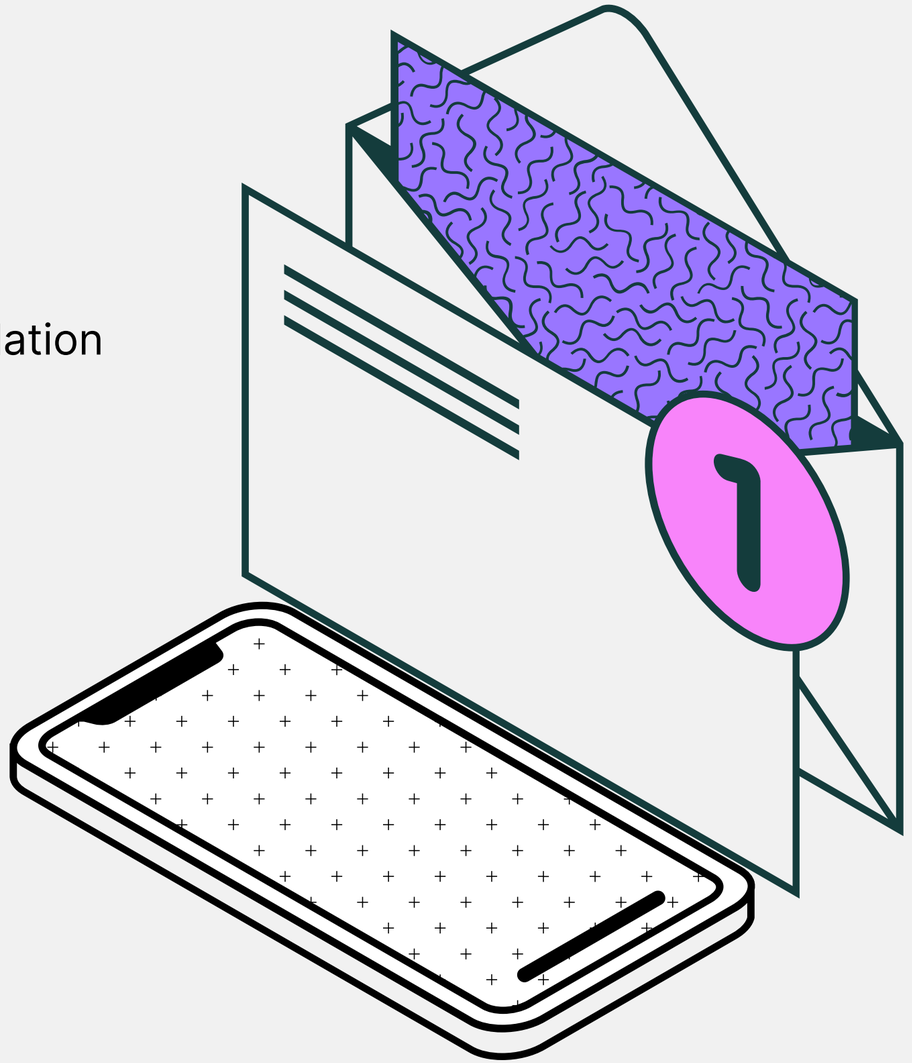
I. Introduction

Short Message Service (SMS) is an essential communication tool in the Philippines.

- 168.3mil cellular mobile connections = 144.5% total population

40% of SMS traffic in the country shows evidence of **manipulated content.**

Spam texts often contain **malicious links, fake promotions, and scams** leading to **financial and identity theft**.



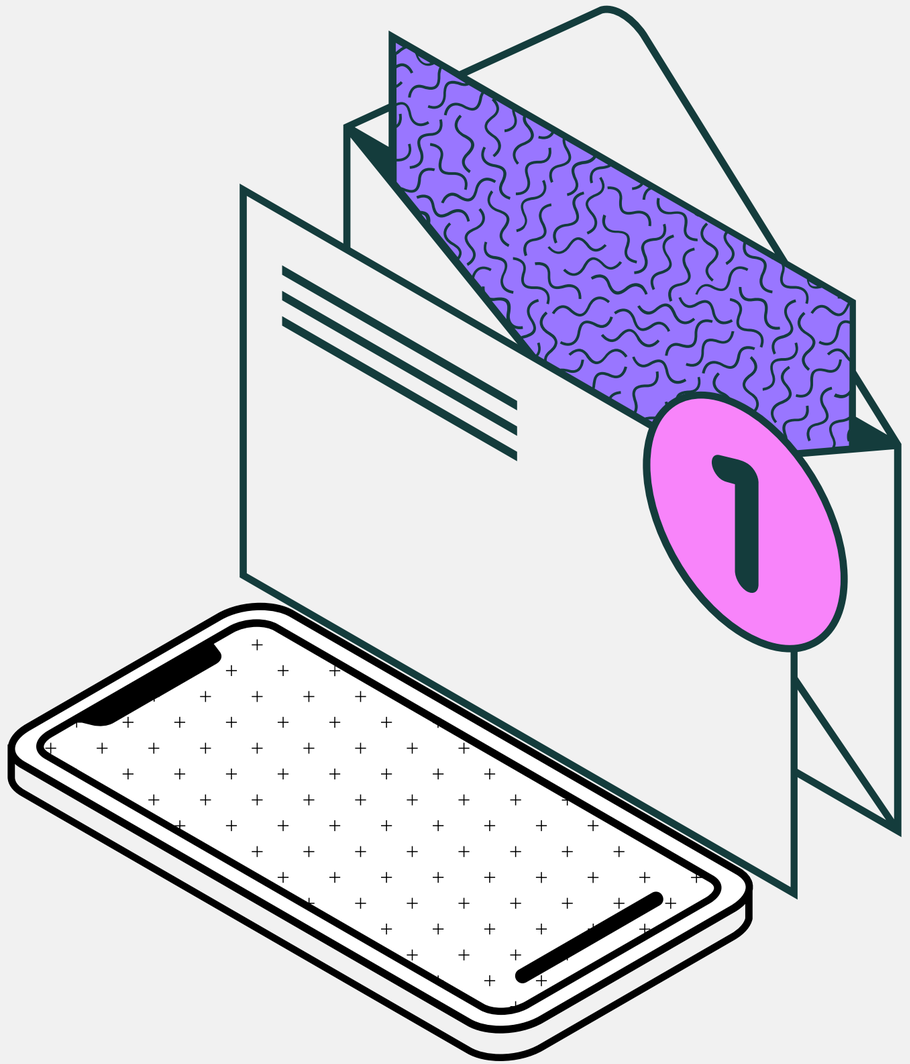
I. Introduction

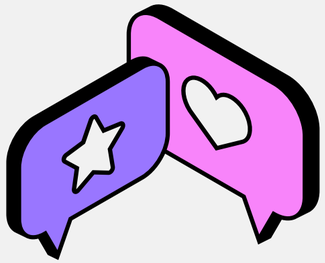
Current Solutions:

Mostly focus on **binary classification**
(spam vs. non-spam)

Need for Multi-Classification:

Spam SMS can be categorized into types like regular, info, ads, and fraud. Not all spam is harmful, but some types are dangerous.

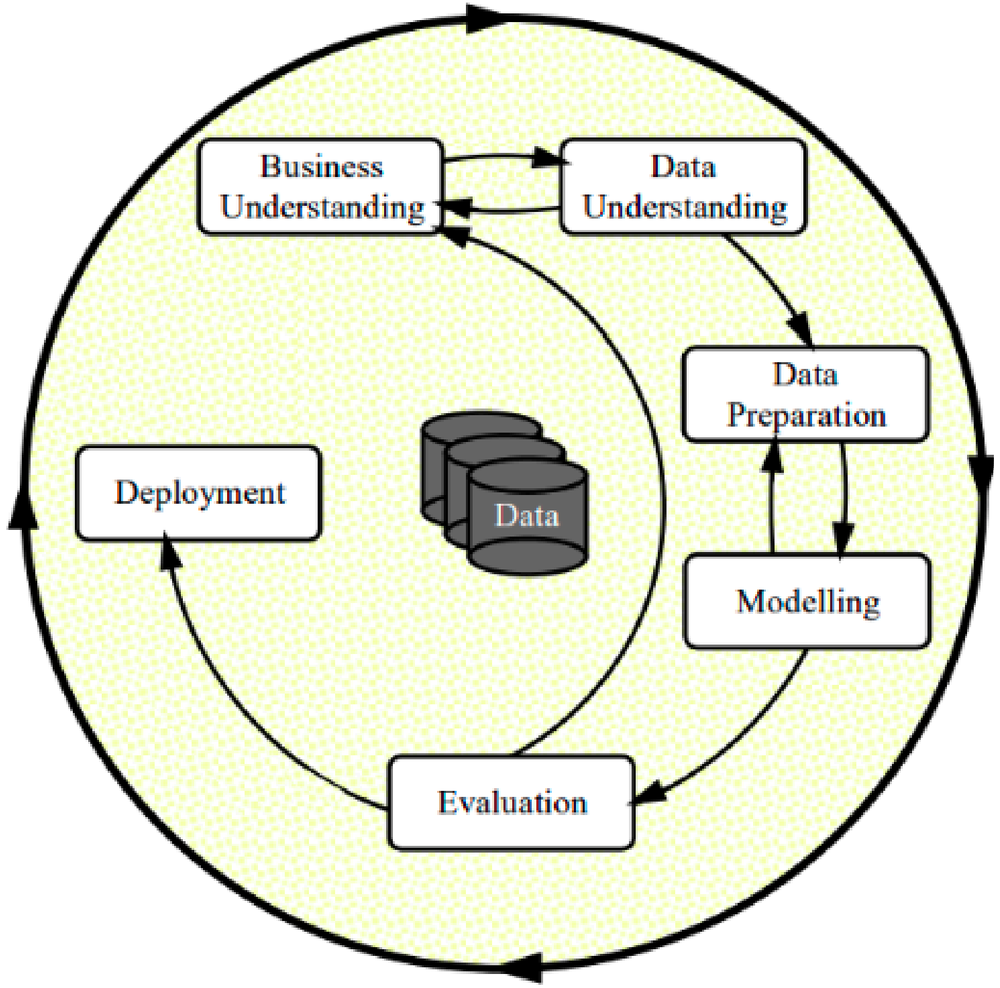




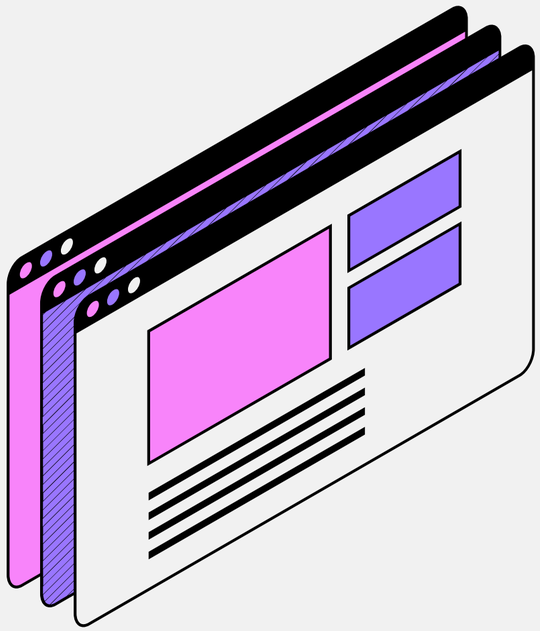
This study explores using unsupervised clustering algorithms**, such as K-means and agglomerative hierarchical clustering, to categorize and analyze spam SMS** in the Philippines.

II. Methodology

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation



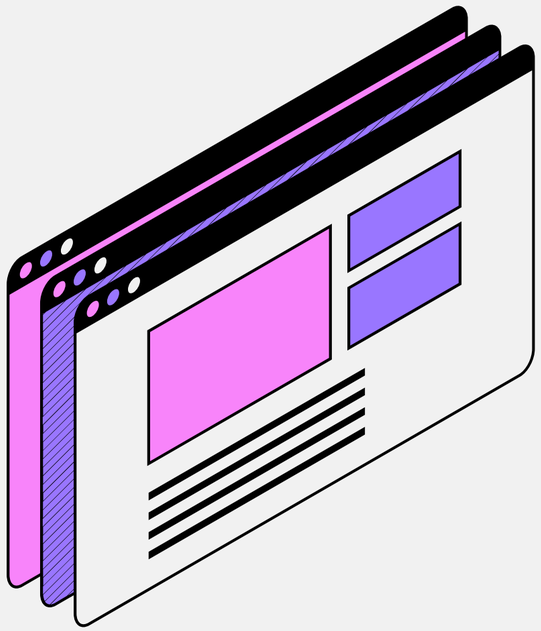
Business Understanding



- **Current Challenges**: While binary classification detects spam, it doesn't capture the diversity of spam types, which vary in risk. Categorizing spam into specific types can provide deeper insights into spammers' intentions.

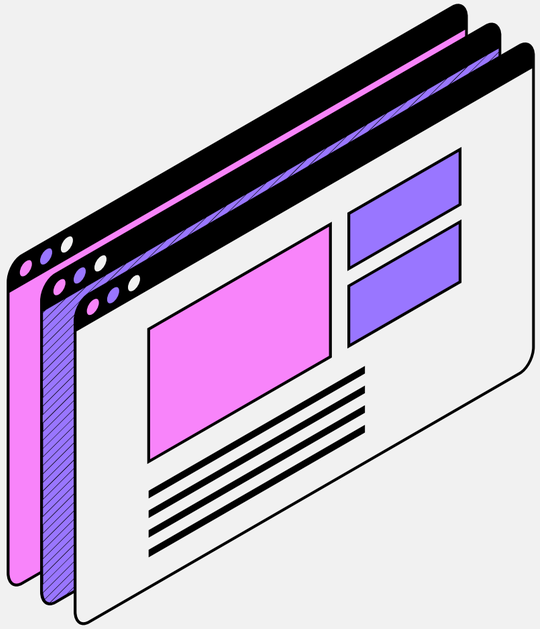
- **Research Focus**: A study used a clustering algorithm to group spam types, creating a labeled dataset for supervised learning with NLP techniques.

Data Understanding



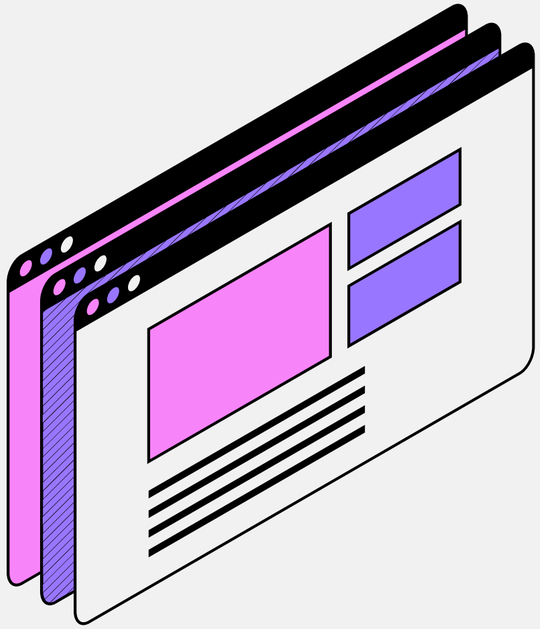
- Acquired from Kaggle: by **‘BwandoWando’**
- Composed of **948 spam SMS** text messages received by the user from November 2022 to November 2024, containing both Filipino and English.
- Texts containing the user’s name were replaced with *<REAL NAME>*
- Text content that is not in string form is marked as *‘Content not supported.’*
- Four variables: *hashed_celphone_number*, ***text***, *date*, and *carrier*

Data Understanding



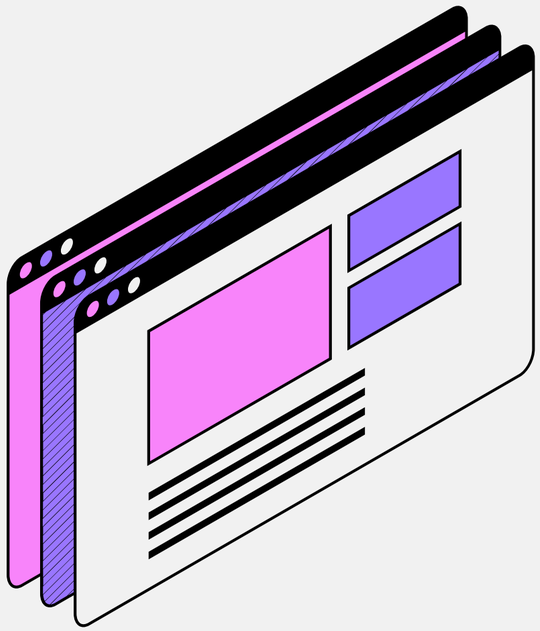
text
betslotsph.tv/KBKGnhr I-double ang halaga ng iyong deposito! 100% deposit bonus para sa unang pagkakataon.
GJPrize.de! Sunday is your chance to win big! Spin the Daily Lucky Wheel for a Samsung Galaxy S23+ 5G (P42,999 value). Newbies get 200% bonus + P3000.
<REAL NAME>, Play with friends and EARN MORE THAN PHP250! Referral BONUS only for you & your close. Click now wpluswow.com
Bagong balita! Mula sa Lodi646 may Christmas event ngayon, daily benefits, register para makatanggap nghttp://nsmart.pro
Welcome ! your have P1222 for S!ot , \nWeb: 11y.life \nGood Luck!C

Data Preparation



- **Text Cleaning:**
  - Dropped non-text columns and converted texts to lowercase.
  - Removed rows with 'content not supported' and replaced <REAL NAME> phrases with empty strings to reduce noise.
  - Removed punctuation and extra spaces for cleaner text.
- **Stopword Removal:** Removed English and Tagalog stopwords using NLTK and Stopwords-ISO.
- **Word Validation:** Invalid words were filtered using NLTK's corpus and a Filipino wordlist.

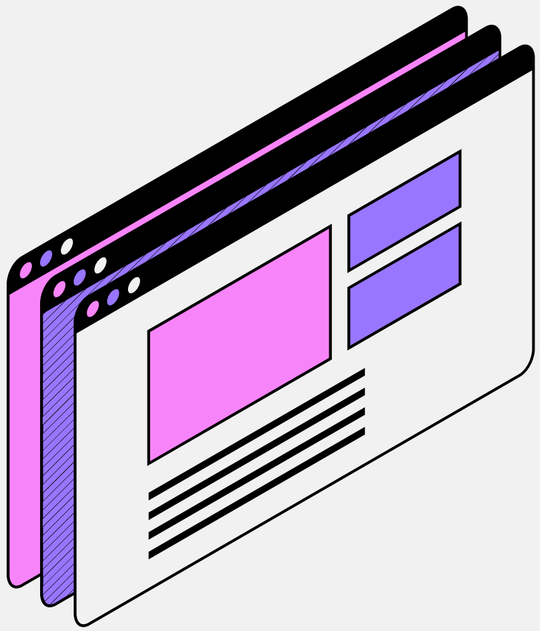
Data Preparation



- **Dataset Split:** The dataset was divided into Tagalog (293 texts) and English (641 texts) datasets after preprocessing.

- **Stemming:** Stemming was not applied due to ambiguity issues, especially with Tagalog, where stemming caused word obfuscation (e.g., **nanalo → alo**). Similar issues were noted in a related study on Spanish and English.

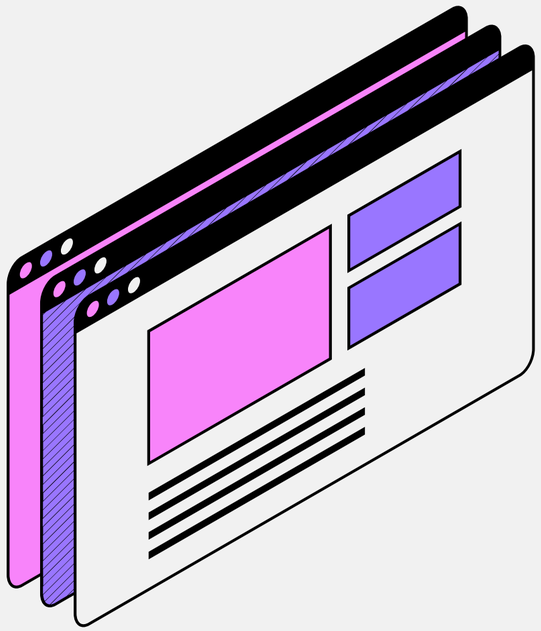
Data Preparation



Tagalog text keywords	English text keywords
magparehistro makakuha tumaya araw makatanggap cash red envelope	sign world get cup benefits receive free seconds
makakuha libre	thank god saturday claim free today new player enjoy deposit bonus win jackpot
bored laro muna free pesos unang deposito register	could let win amount okada even
sali photo contest promotion upang manalo manghang premyo	good day gcash please advised asking everyone verify account avoid deactivation tomorrow
gawin unang deposito makakuha deposit bonus doblehin itong palampasin	x join us free today auto get extra deposit trusted website

TABLE II. RANDOM TEXT SAMPLES AFTER PREPROCESSING

Modelling



- **Clustering Techniques:**

  Hierarchical Clustering:

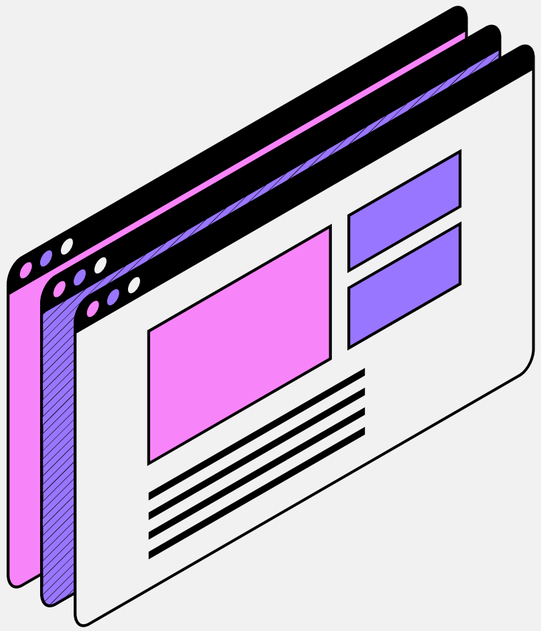
  - Used Agglomerative Clustering with Ward's linkage to create a dendrogram.

  K-Means Clustering:

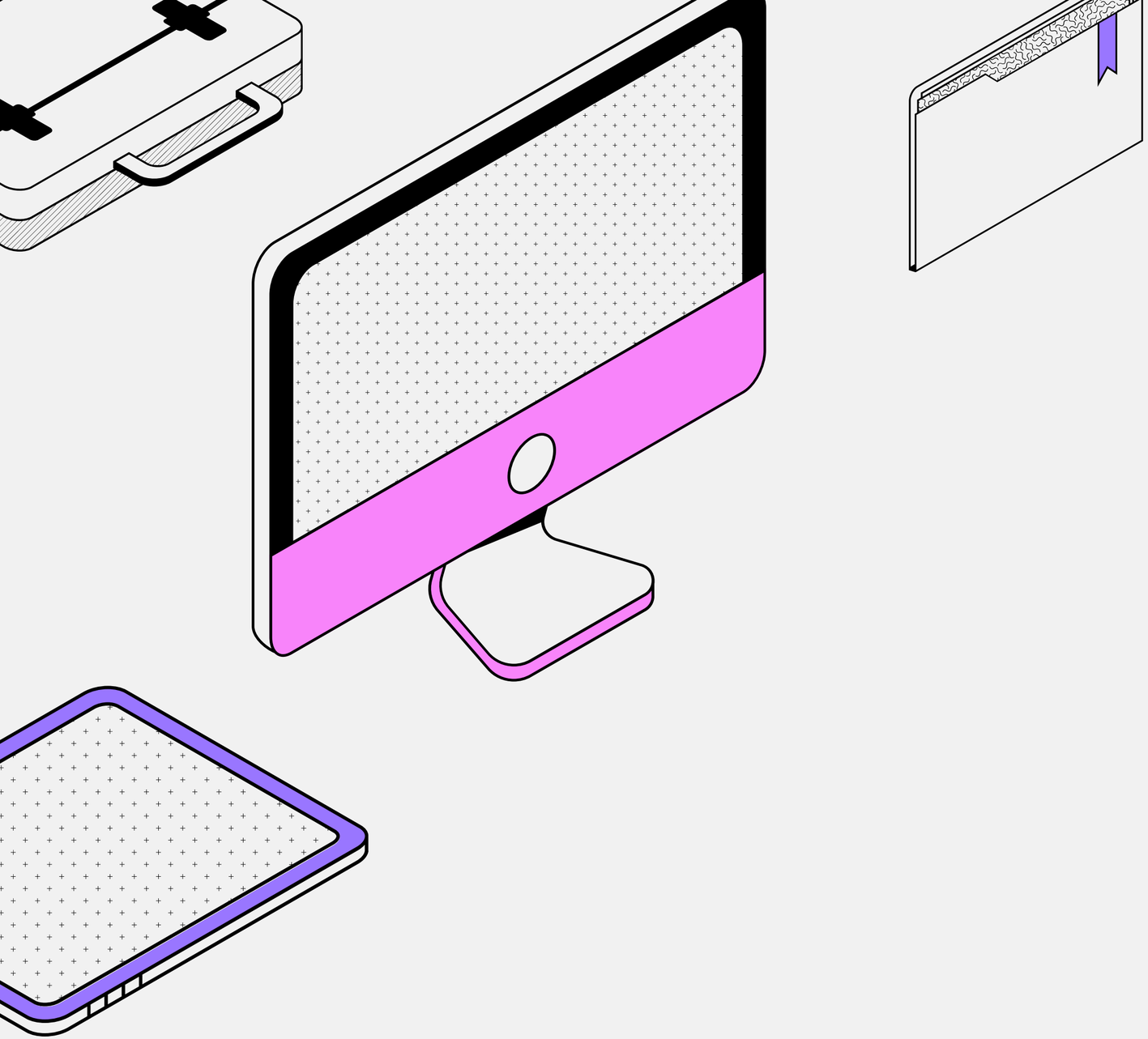
  - The optimal number of clusters (k) was determined using the Elbow method and silhouette scores.

- **Data Processing:** Texts were vectorized using Bag of Words (**BOW**) and reduced to **60 dimensions** via PCA for efficient clustering.

Evaluation

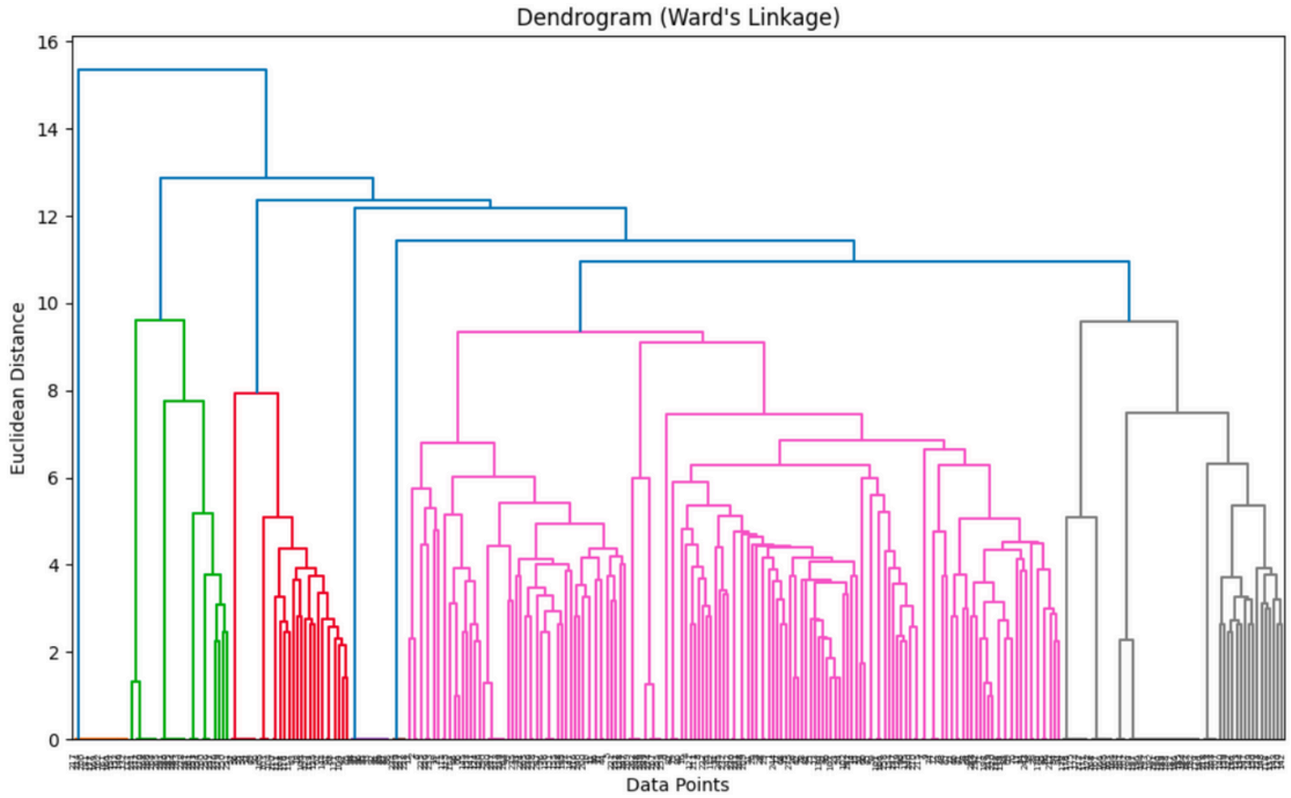


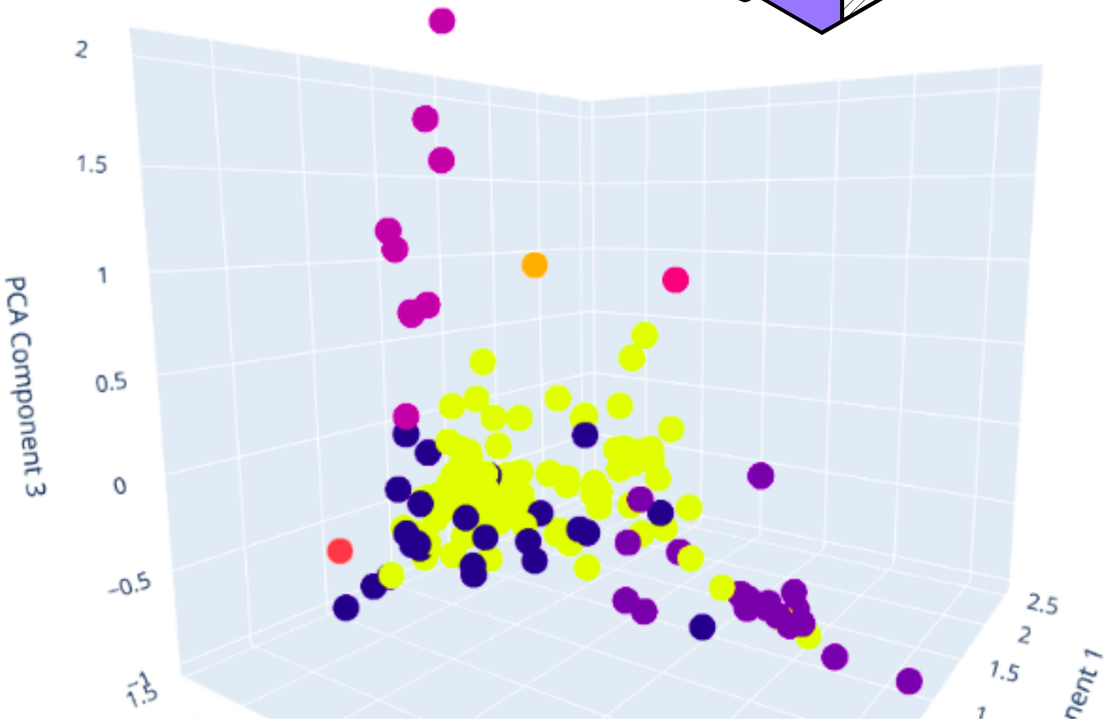
- **Quantitative Metrics:**
  - **Silhouette Score:** Measures cluster separation; higher values indicate better-defined clusters.
  - **Davies-Bouldin Index:** Compares within-cluster similarity to nearest clusters; lower values indicate compact, distinct clusters.
  - **Calinski-Harabasz Score:** Evaluates within-cluster vs. between-cluster dispersion; higher values signify well-defined clusters.
- **Qualitative Analysis:**
  - **Manual inspection** of clusters to assess meaningful and consistent grouping.
  - **Visualization:** 3D scatter plots and dendrograms were used to visually evaluate cluster separation and structure.



Results and Discussion

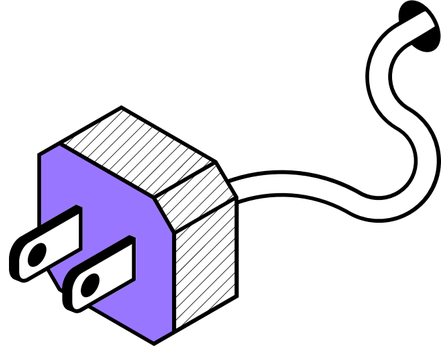
Hierarchical Clustering - Tagalog Dataset





- Silhouette Score: **0.1812**
- Davies-Bouldin Index (DBI): **1.9063**
- Calinski-Harabasz (CH) Score: **20.1140**

Hierarchical Clustering - Tagalog Dataset

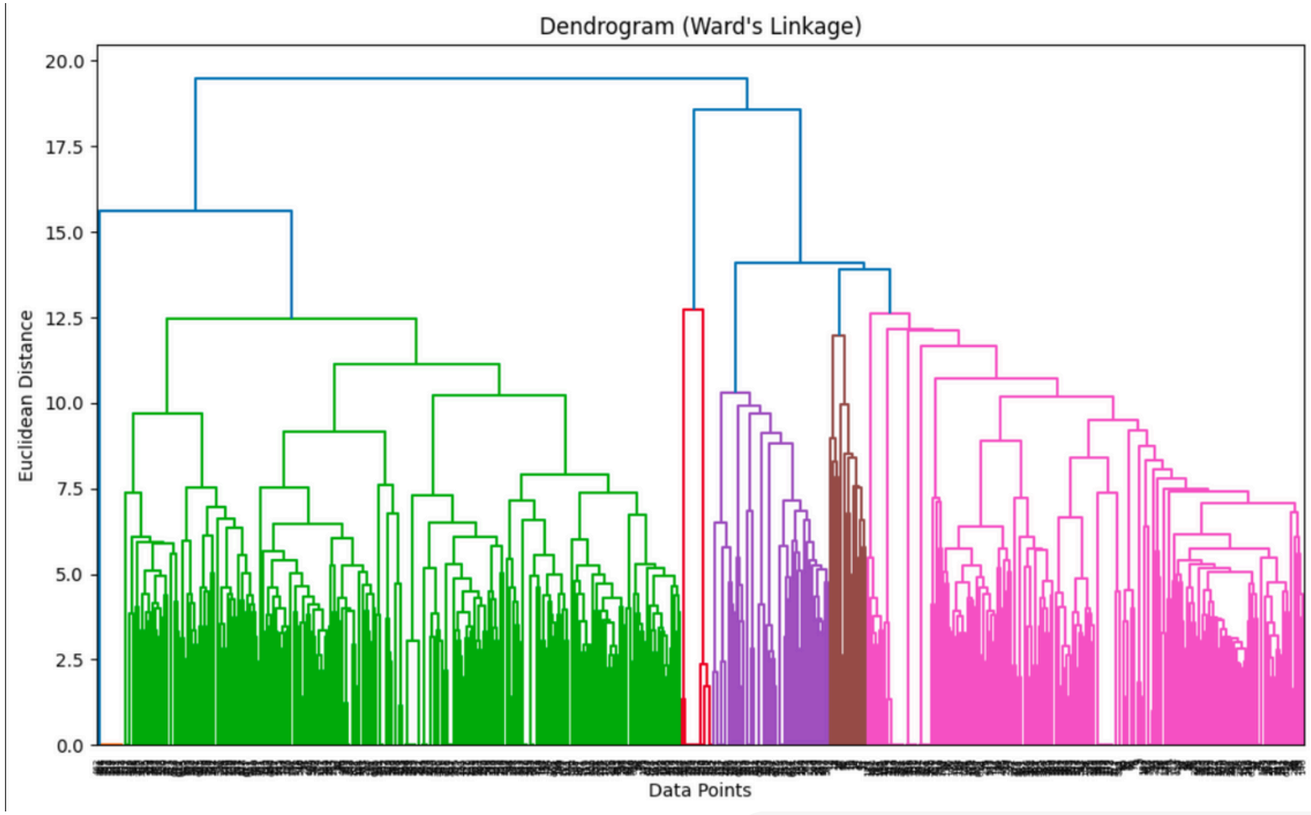


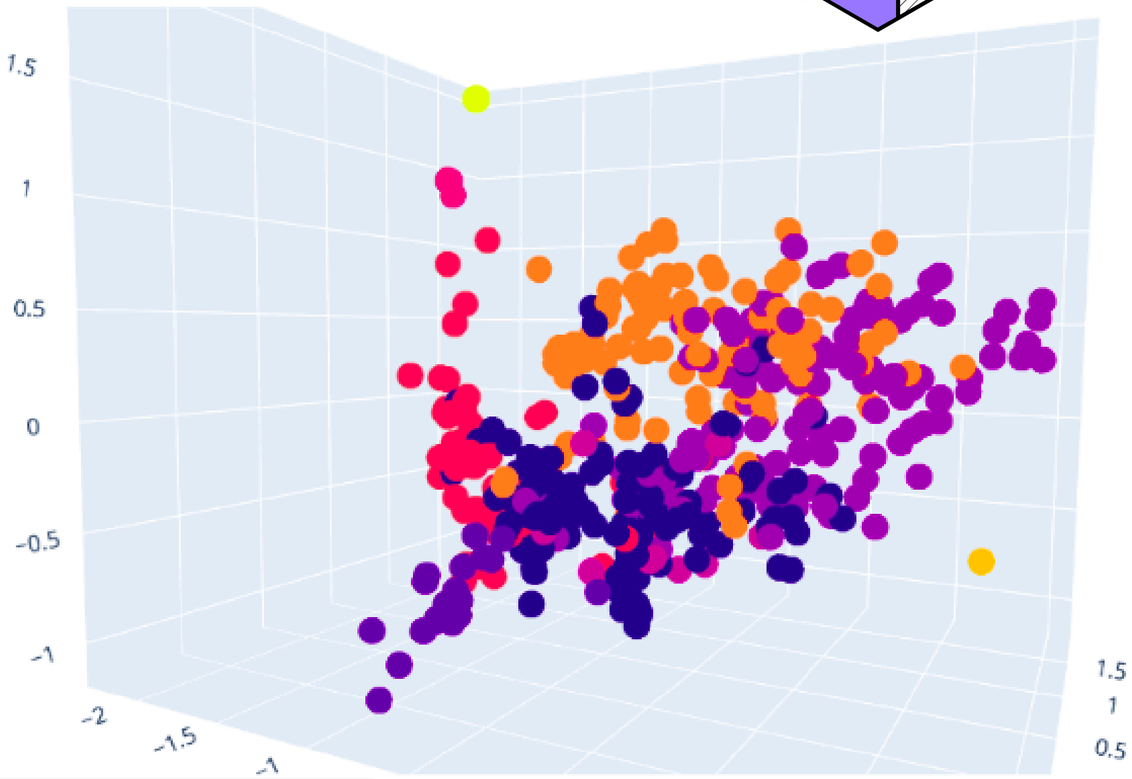
Cluster	Description	Size	Percent
0	libre, magparehistro, taya, pera	69	23.55%
1	encouraging deposits (all)	29	9.90%
2	claims, nanalo, na/maka-tanggap, makuha, credit— received money scams	24	8.19%
3	sumali, bonus, deposit	14	4.78%
4	kumita, Youtube	10	3.41%
5	goodluck, swerte, bet	4	1.37%
6	others (bonus, makuha, araw)	143	48.81%

Cluster	Description	Points/Size	Percent
1 (0, 5)	libre stuff, register, taya	73	24.92%
2 (1, 3)	encouraging deposits (all)	43	14.68%
3	claims, nanalo, na/maka-tanggap, makuha, credit— received money scams	24	8.19%
4	kumita, Youtube	10	3.41%
5	others (bonus, makuha, araw)	143	48.81%

Clustering Results After Merging (TD)

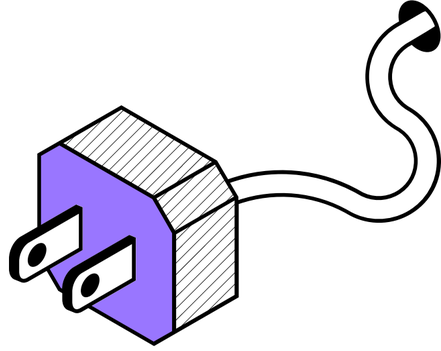
Hierarchical Clustering - English Dataset





- Silhouette Score: 0.0889
- Davies-Bouldin Index: 2.3721
- Calinski-Harabasz Score: 24.4588

Hierarchical Clustering - English Dataset

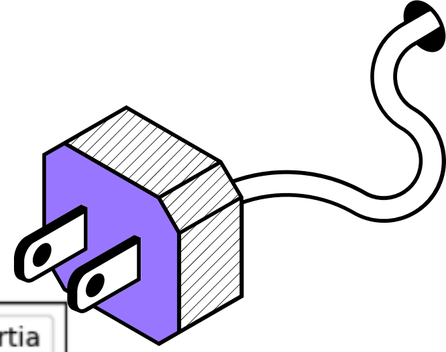


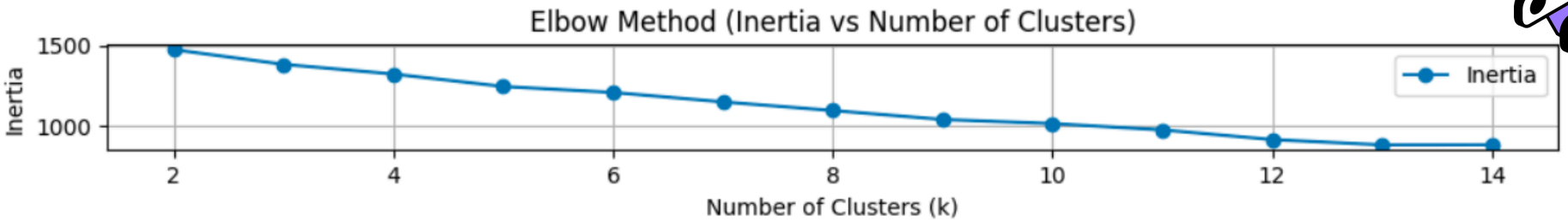
Cluster	Description	Size	Percent
0	Mostly promotional residential sales, some credit cards and bank scams	28	4.37%
1	Mixed - ewallet, gambling, deposit, ads, bonus, congratulations	193	30.11%
2	Win, register - some rouge ewallet, gov phishing, gambling	186	29.02%
3	Roulette, game - gambling, raffles	9	1.40%
4	Banco de oro scams	7	1.09%
5	Bank, tracking, ewallet scams with some rouge ads	61	9.52%
6	Register, deposit, bonus - gambling, raffles	133	20.75%
7	Free rewards login— gambling	14	2.18%
8	Banco de oro scams	10	1.56%

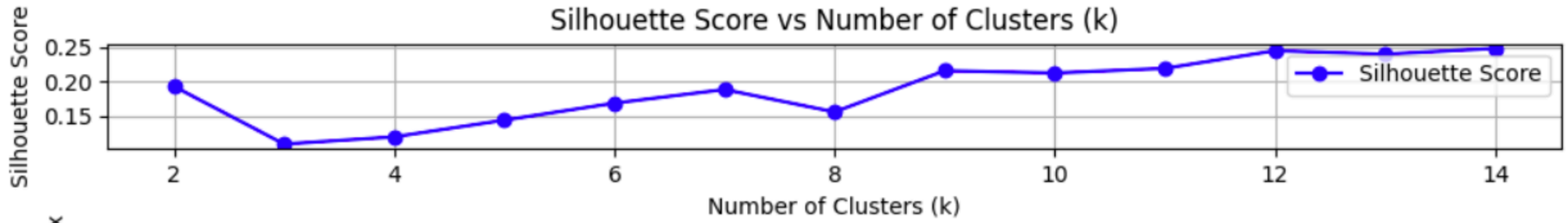
Cluster	Description	Size	Percent
1	Promotional residential sales, credit cards, bank	28	4.37%
2	Mixed - ewallet, gambling, deposit, ads, bonus, congratulations	193	30.11%
3 (2, 3, 6, 7)	Win, register - some rouge ewallet, gov phishing, gambling	342	53.35%
4, 5, 8	Banco de oro scams	78	1.09%
5	others (bonus, makuha, araw)	143	48.81%

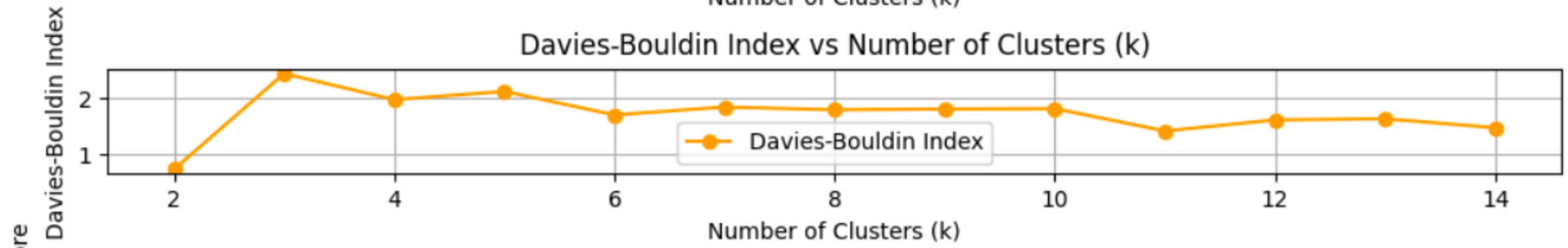
Clustering Results After Merging (TD)

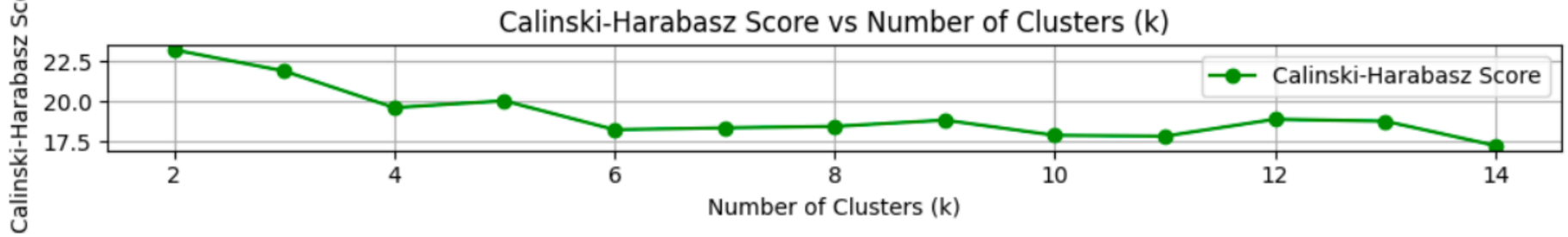
K-Means Clustering



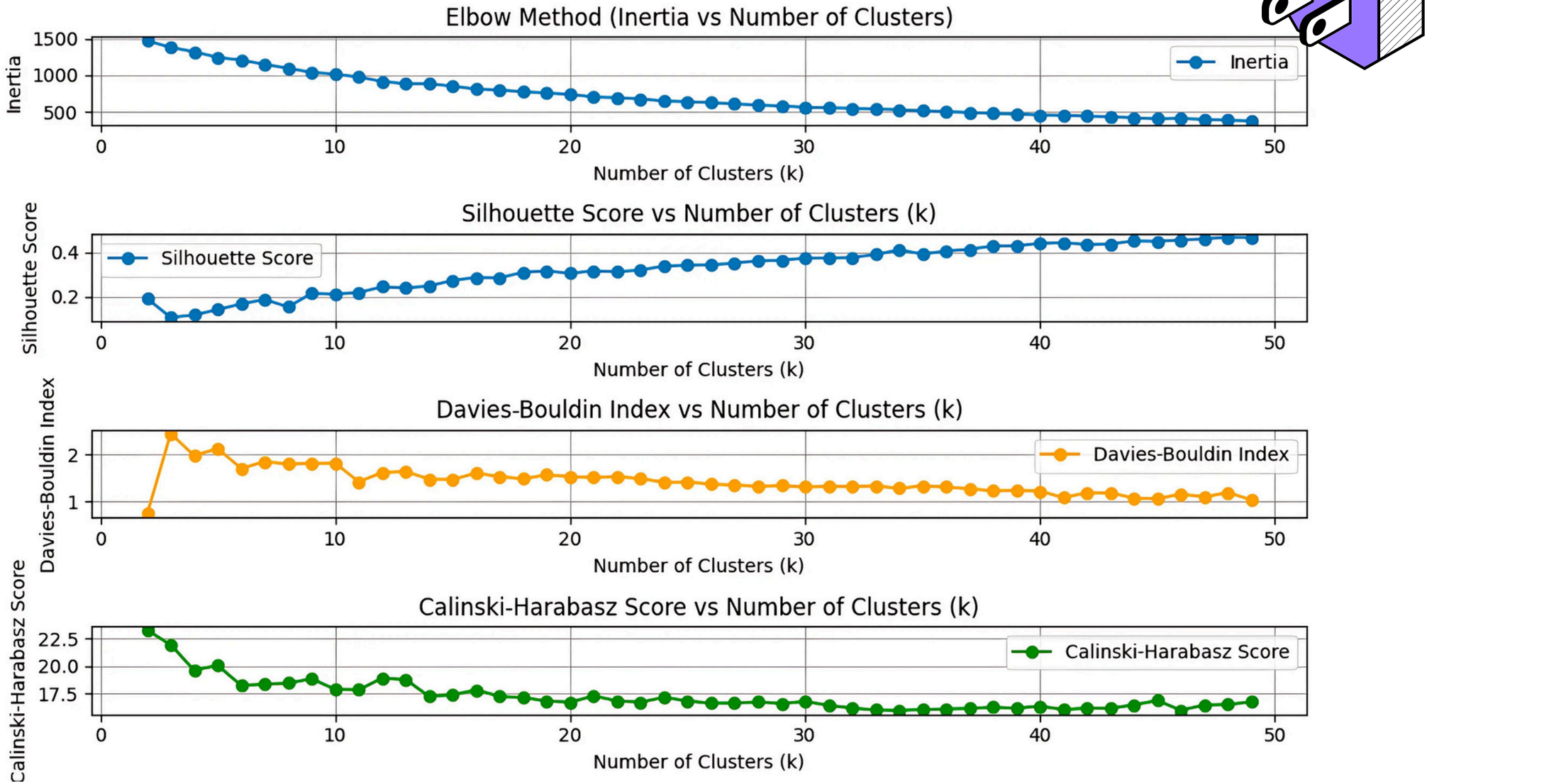




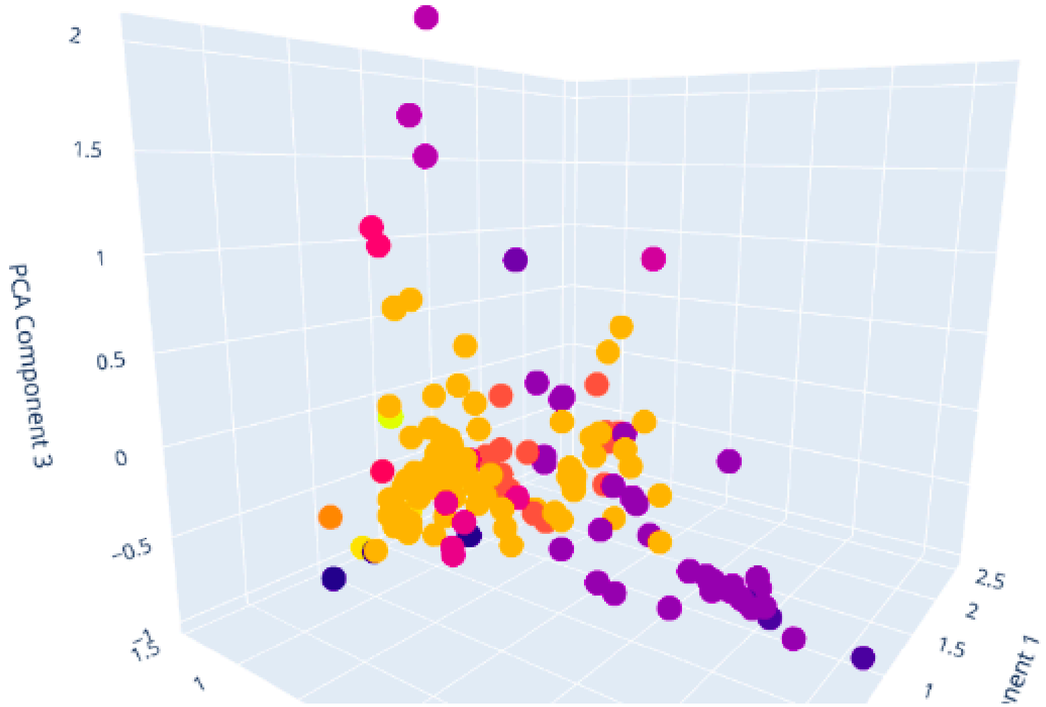




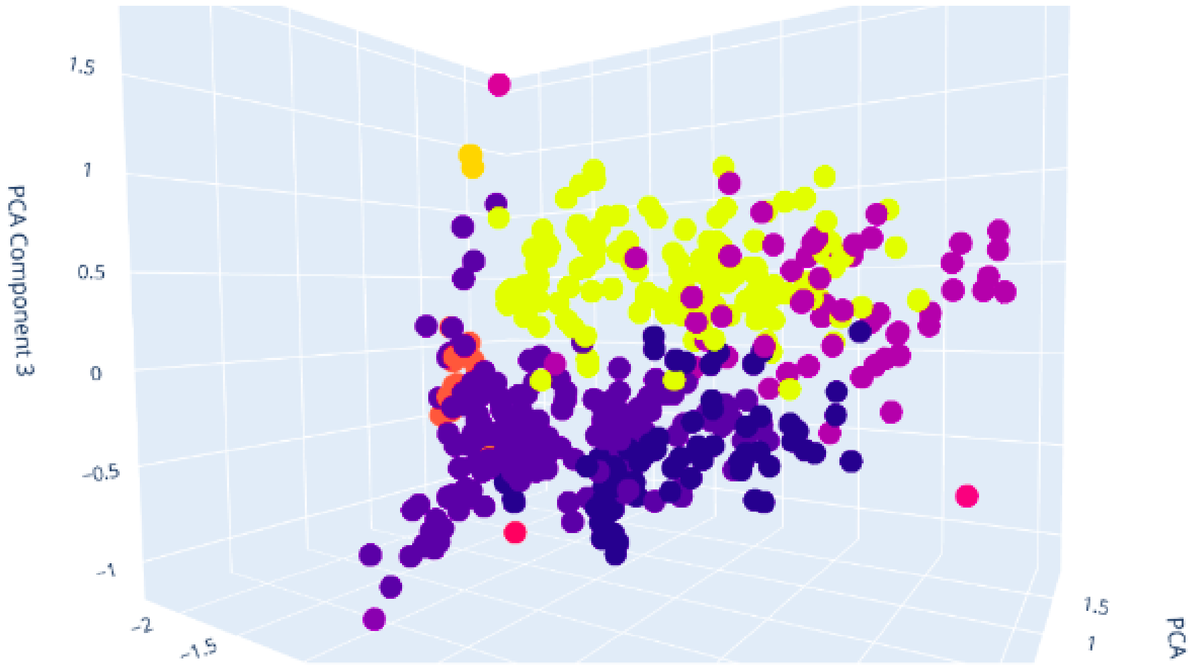
K-Means Clustering



K-Means Clustering



3D Visualization of Tagalog Dataset using KMS, k = 14

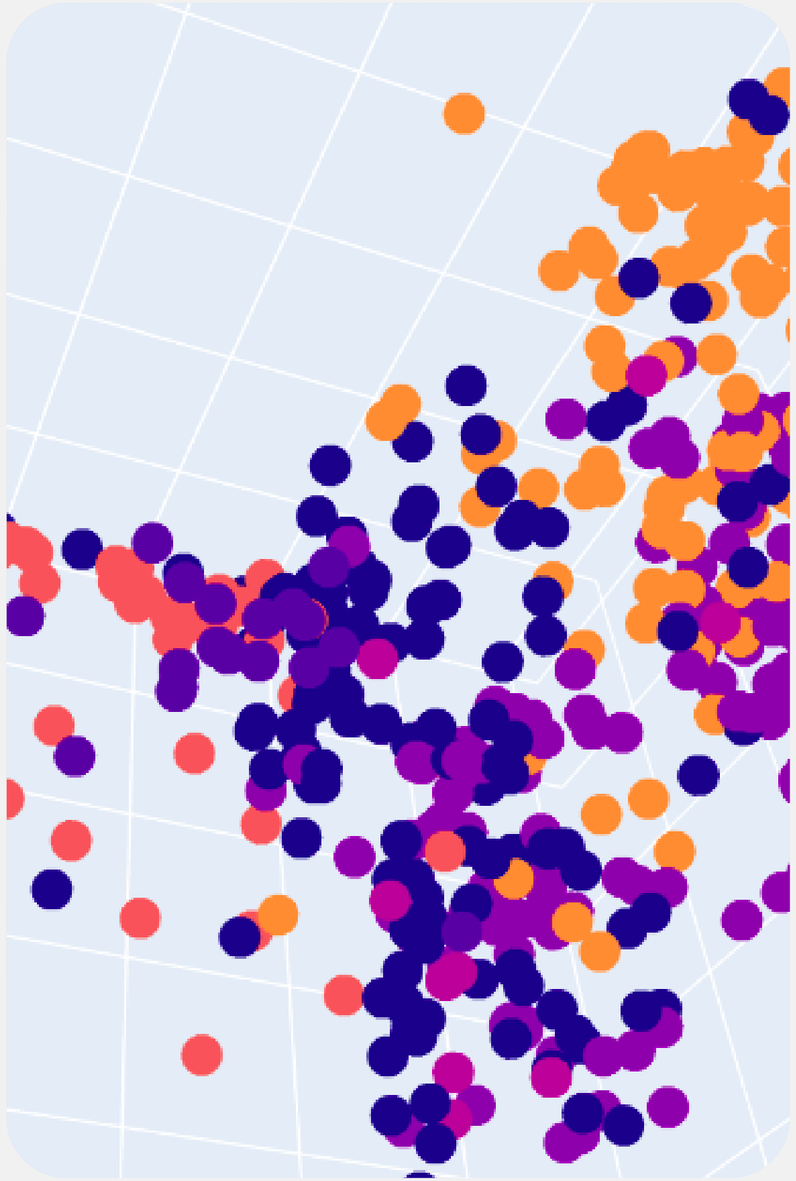


3D Visualization of English Dataset using KMS, k = 11

Conclusion

Insights from Clustering:

- **Hierarchical Clustering** identifies broad spam categories, such as phishing patterns in Tagalog data and promotional content (e.g., real estate ads) in English data. However, some clusters require manual adjustment for accuracy.
- **K-Means Clustering** struggled with clear categorization due to challenges in determining the optimal cluster count. While higher clusters (e.g., 50) effectively group similar messages, thematic clarity is compromised.



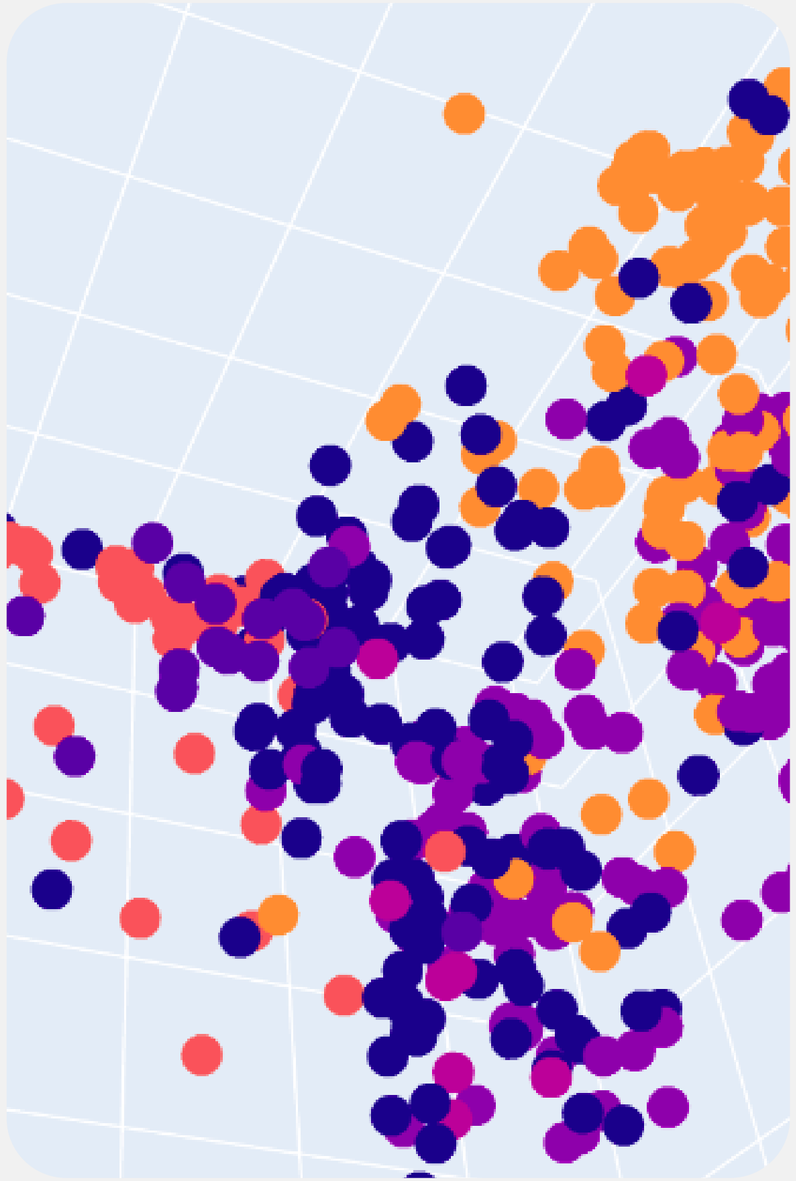
Conclusion

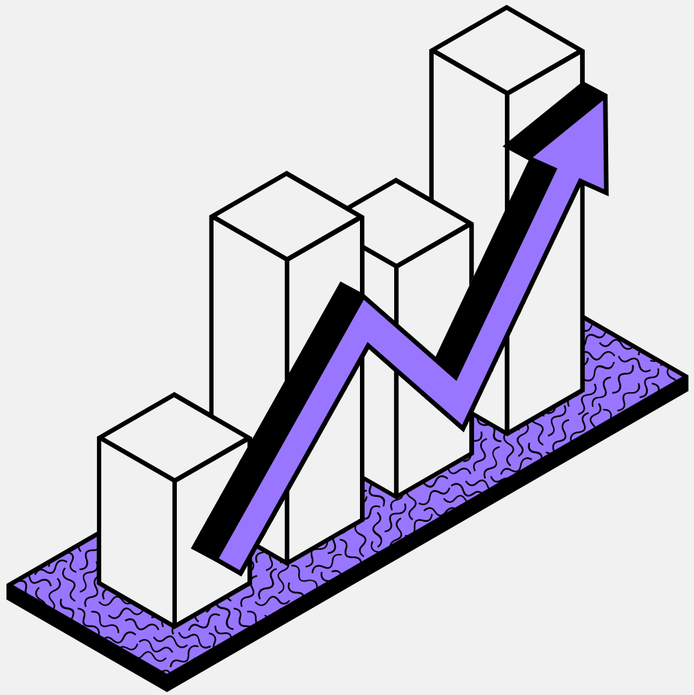
Applications:

- Labeling large spam datasets:
  - Hierarchical Clustering: Provides initial structure for human refinement.
  - K-Means: Groups template-like, repetitive messages.
- Supports dataset cleaning and organization for machine learning projects.

Recommendations for Improvement:

- Incorporate NLP techniques to extract linguistic features.
- Expand the dataset beyond fraud and promotions for better categorization.
- Include human review to refine clusters for greater accuracy.





Overall Potential

Both clustering methods offer value for organizing spam datasets, with potential applications in labeling, identifying spam templates, and improving spam detection systems through enhanced preprocessing and analysis.

Thank you.