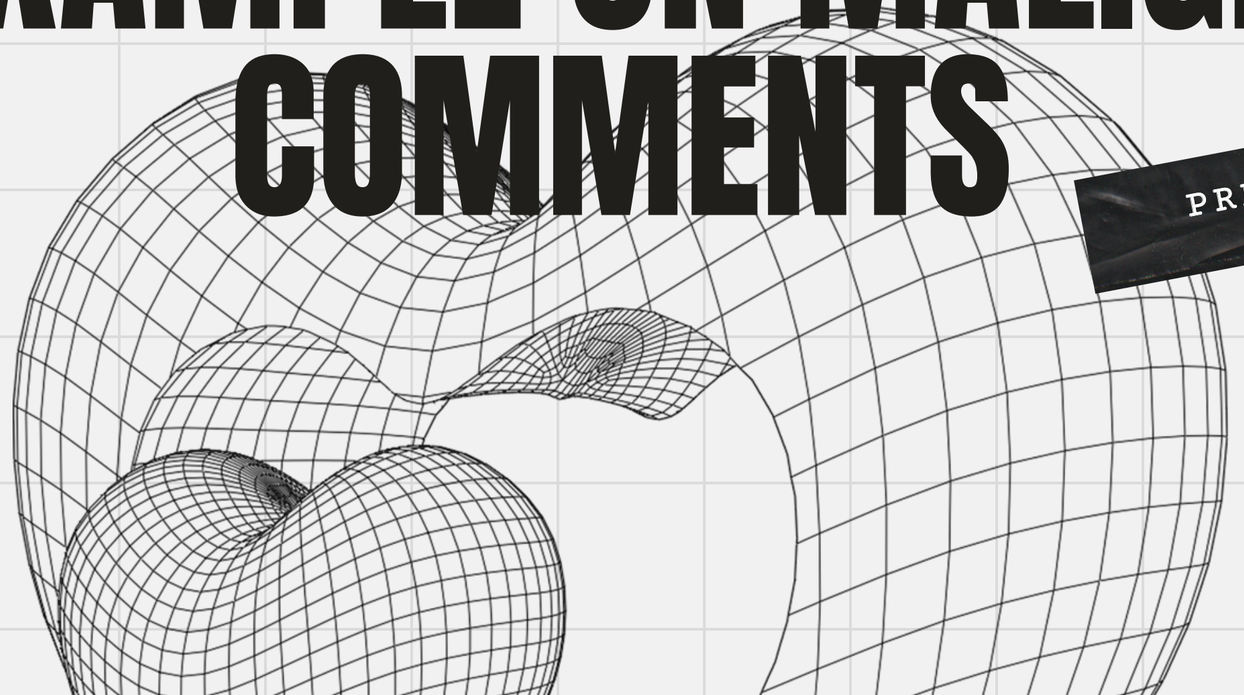
CERVINI STELLA, SABINO GIUSEPPE

2023/2024

TM&S PROJECT

CLASSIFICATION VS CLUSTERING: AN EXAMPLE ON MALIGNANT COMMENTS

PRESENTATION



TM&S PROJECT

CERVINI STELLA, SABINO GIUSEPPE

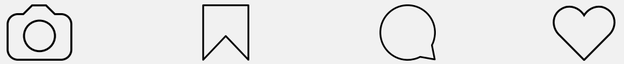
TABLE OF CONTENT

CERVINI STELLA, SABINO GIUSEPPE

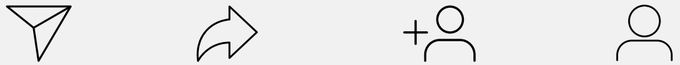
INTRODUCTION

The Internet, once praised for its open communication and limitless knowledge, has paradoxically turned into a breeding ground for malicious comments, a poisonous undercurrent threatening to contaminate the very essence of online discourse.

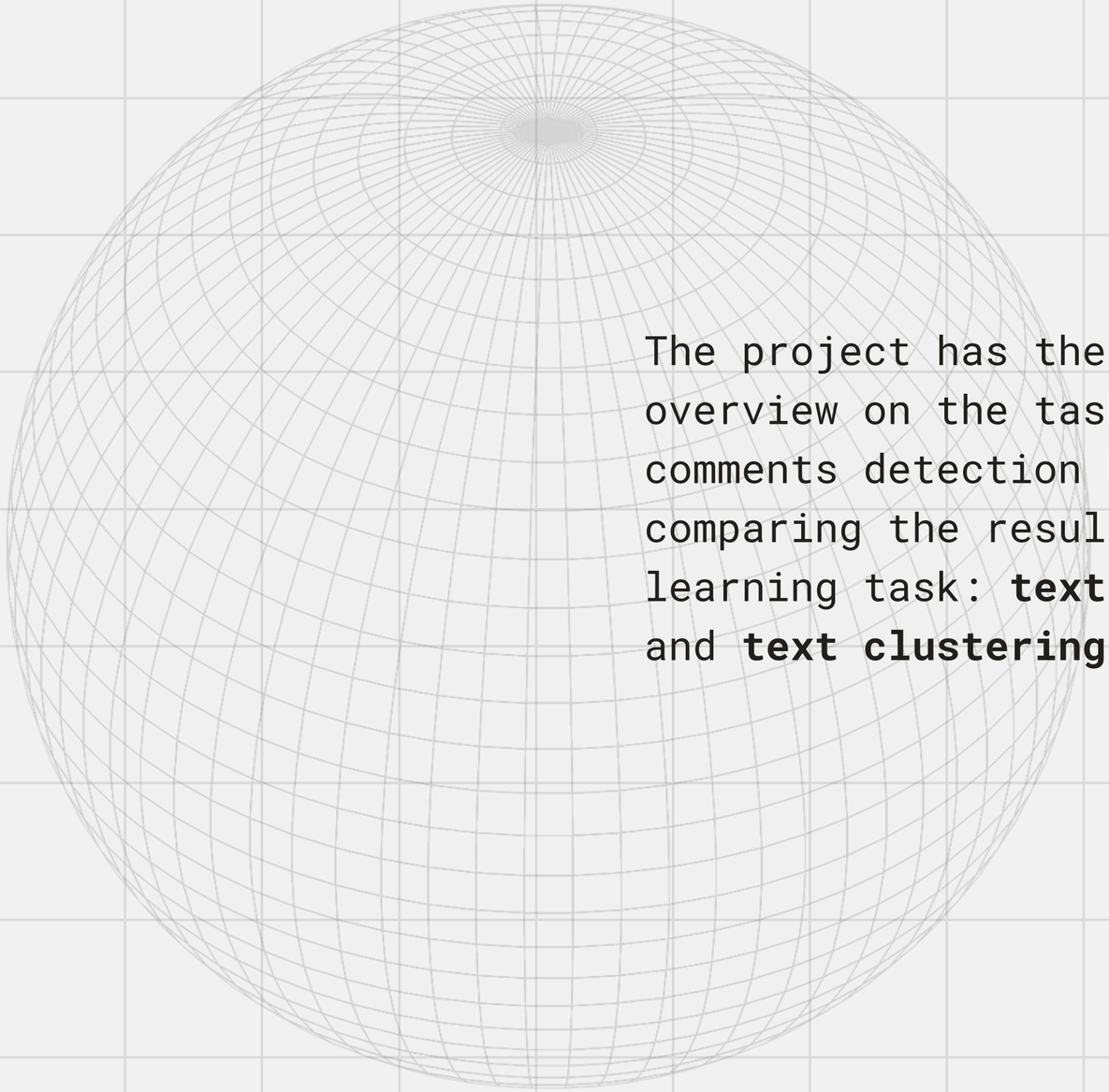
THESE MALICIOUS MESSAGES CAN HAVE SEVERE CONSEQUENCES, CAUSING EMOTIONAL DISTRESS, SILENCING MARGINALIZED VOICES, AND EVEN INCITING REAL-WORLD VIOLENCE AND THEIR IMPACT EXTENDS BEYOND INDIVIDUAL EXPERIENCES, PERMEATING THE BROADER ONLINE ECOSYSTEM.



DEVELOPING AND IMPLEMENTING EFFECTIVE AUTOMATED CLASSIFICATION SYSTEMS IS A COMPLEX ENDEAVOR. THE DIVERSE NATURE OF MALICIOUS LANGUAGE, ENCOMPASSING INSULTS AND HATE SPEECH POSES A SIGNIFICANT CHALLENGE. ADDITIONALLY, THE INHERENT AMBIGUITY AND CONTEXT-DEPENDENT NATURE OF NATURAL LANGUAGE FURTHER COMPLICATE THE TASK.



PROJECT GOALS



The project has the goal to provide an overview on the task of malignant comments detection by exploiting and comparing the results of two machine learning task: **text classification** and **text clustering**.

CERVINI STELLA, SABINO GIUSEPPE

CLASSIFICATION VS CLUSTERING

While classification and clustering are distinct machine learning techniques, they share a common thread: their ability to extract patterns and insights from data. Both techniques seek to understand the data's structure, but they approach this task from different angles.

CLASSIFICATION ASSIGNS DATA POINTS TO PREDEFINED CATEGORIES. THIS TECHNIQUE IS OFTEN USED FOR *SUPERVISED LEARNING*, WHERE THE ALGORITHM IS TRAINED ON LABELED DATA.
BY LEARNING FROM LABELED DATA, THE ALGORITHM CAN IDENTIFY THE PATTERNS THAT DISTINGUISH BETWEEN THE DIFFERENT CATEGORIES.
ONCE TRAINED, THE ALGORITHM CAN BE USED TO CLASSIFY NEW DATA POINTS.

5K 3K 9K

CLUSTERING GROUPS DATA POINTS TOGETHER BASED ON THEIR SIMILARITY.
THIS TECHNIQUE IS OFTEN USED FOR *UNSUPERVISED LEARNING*, WHERE THE ALGORITHM IS NOT GIVEN LABELED DATA. INSTEAD, THE ALGORITHM MUST DISCOVER THE PATTERNS IN THE DATA ON ITS OWN AND GROUP THE DATA POINTS ACCORDINGLY.

3K 2K 8K

DATA

The data was collected from Kaggle. The data is composed by two datasets: a training set, which has approximately 159,000 samples and the test set which contains nearly 153,000 instances. All the data samples contain 8 fields.

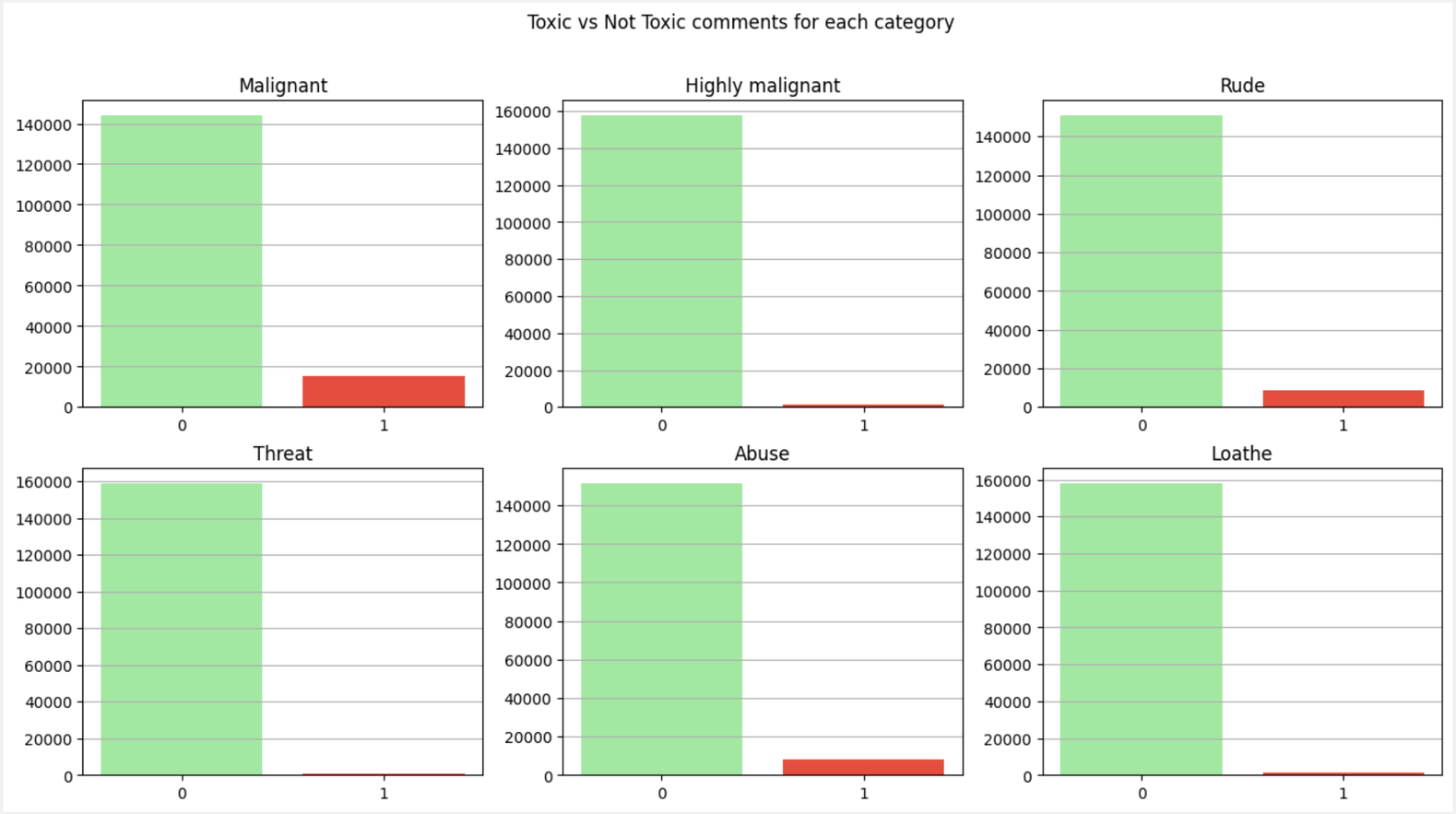
- *ID*: IT INCLUDES UNIQUE IDS ASSOCIATED WITH EACH COMMENT TEXT GIVEN.
- *COMMENT TEXT*: THIS COLUMN CONTAINS THE COMMENTS EXTRACTED FROM VARIOUS SOCIAL MEDIA PLATFORMS

- *MALIGNANT*: IT IS THE LABEL COLUMN, WHICH INCLUDES VALUES 0 AND 1, DENOTING IF THE COMMENT IS MALIGNANT OR NOT.
- *HIGHLY MALIGNANT*: IT DENOTES COMMENTS THAT ARE HIGHLY MALIGNANT AND HURTFUL.
- *RUDE*: IT DENOTES COMMENTS THAT ARE VERY RUDE AND OFFENSIVE.
- *THREAT*: IT CONTAINS INDICATION OF THE COMMENTS THAT ARE GIVING ANY THREAT TO SOMEONE.
- *ABUSE*: IT IS FOR COMMENTS THAT ARE ABUSIVE IN NATURE.
- *LOATHE*: IT DESCRIBES THE COMMENTS WHICH ARE HATEFUL AND LOATHING IN NATURE.

THE LABEL CAN BE EITHER 0 OR 1, WHERE DENOTES 0 A NO WHILE 1 DENOTES A YES.
THERE ARE VARIOUS COMMENTS WHICH HAVE MULTIPLE LABELS.

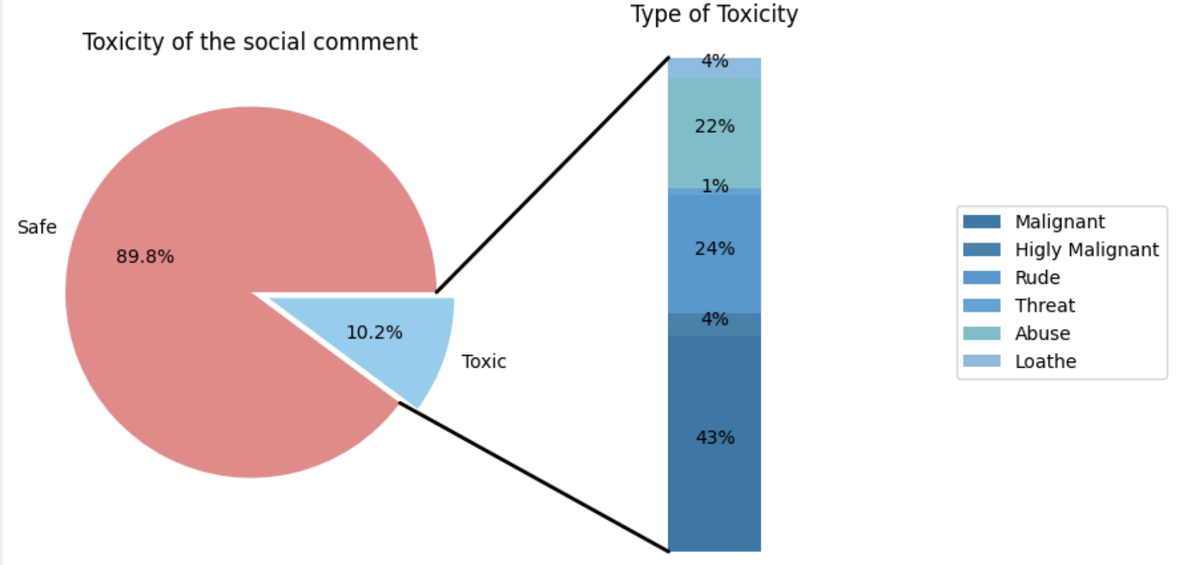
CERVINI STELLA, SABINO GIUSEPPE

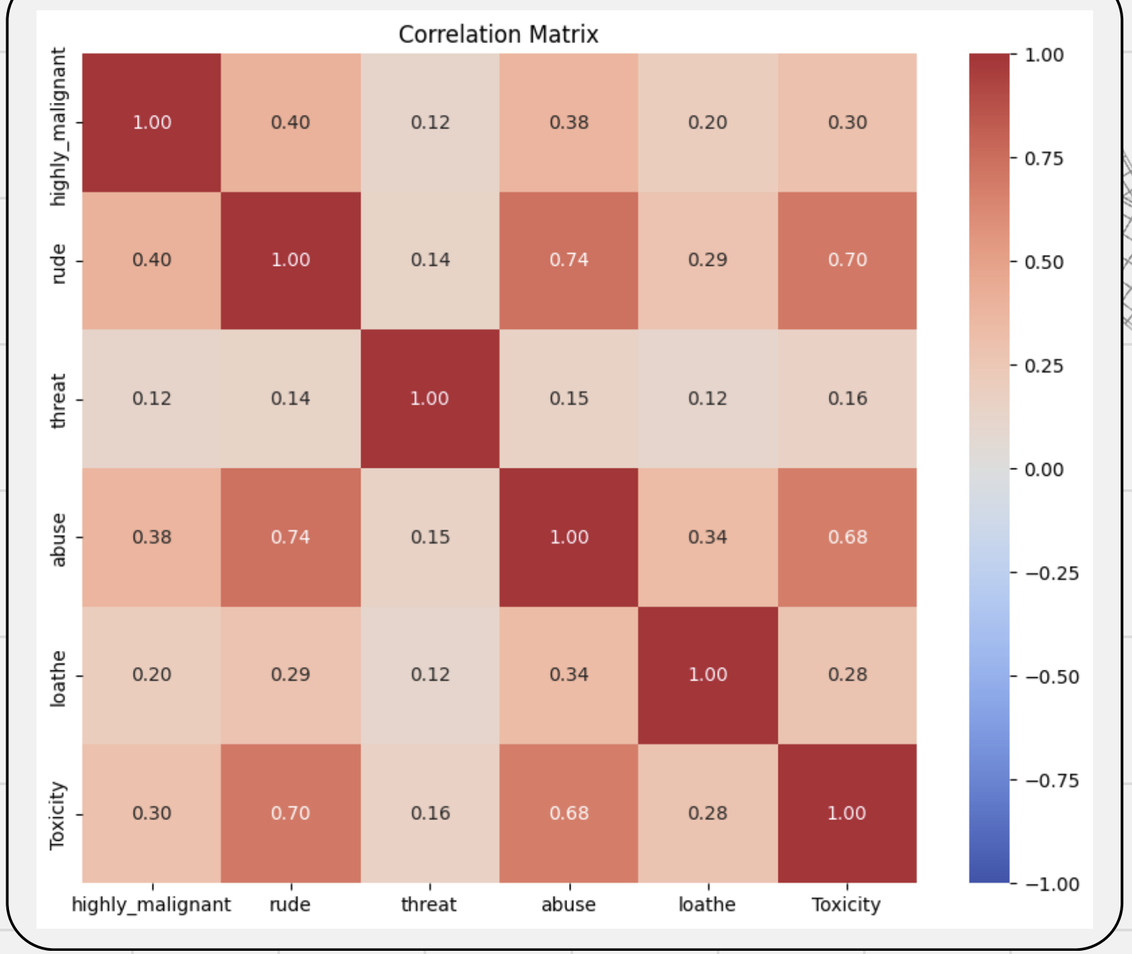
DATA EXPLORATION 1/2



CERVINI STELLA, SABINO GIUSEPPE

DATA EXPLORATION 2/2





STEP 1

Remove URLs

STEP 2

Replace Newlines

STEP 3

Convert to Lowercase

STEP 4

Strip Whitespace

TEXT PREPROCESSING

The aim of the text preprocessing phase is to clean and preprocess textual data, making it more suitable for natural language processing (NLP) tasks such as sentiment analysis or classification.

CERVINI STELLA, SABINO GIUSEPPE

STEP 5

Tokenization

STEP 6

Remove Stopwords

STEP 7

Remove Special Characters

STEP 8

Remove Long Tokens

TEXT PREPROCESSING

Preprocessing is likely to lead to improved accuracy, reduced dimensionality, and improved interpretability of the data. It is also important because it can help to reduce the dimensionality of the data and improve data interpretability.

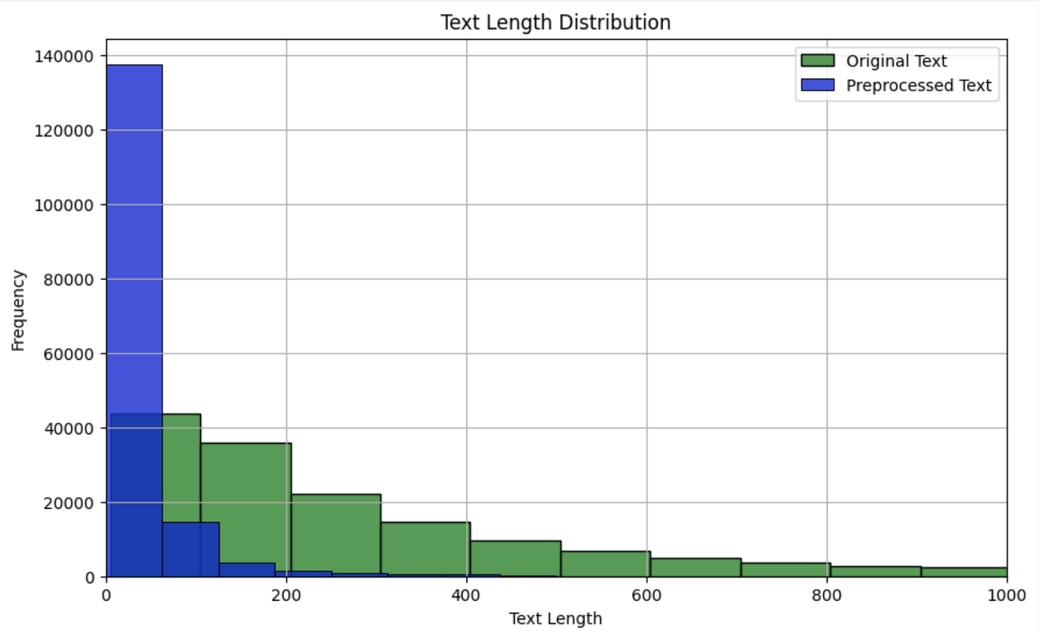
CERVINI STELLA, SABINO GIUSEPPE

STEP 9

Lemmatization

STEP 10

Remove Empty Tokens



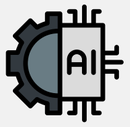
TEXT PREPROCESSING

TM&S PROJECT

09

TEXT CLASSIFICATION

Binary classification to predict to which of the predefined classes (Malignant and Not Malignant) the comments belongs.



MACHINE LEARNING

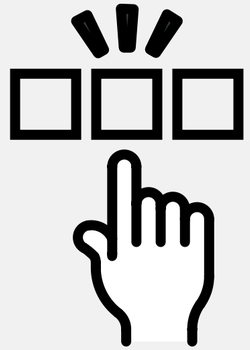
The use of four models is planned:
Logistic regression, Random forest, XGBoost (xgb), and Adaboost (ada)

DEEP LEARNING

ElectraSmall model to improve the result

RESULT EVALUATION

Metrics to compare and evaluate models



MODEL SELECTION

Evaluation of different models is conducted to select the one demonstrating superior Recall for the final test. Identify a model showcasing the best predictive capability on the validation dataset to ensure an accurate model during the testing phase.

- Parameters optimization
- TFidfVectorizer

CERVINI STELLA, SABINO GIUSEPPE

ACCURACY METRICS

The metrics are averaged between the two class and the accuracy is considered on validation set.

METRICS	LOGISTIC REGRESSION	RANDOM FOREST	XGBOOST	ADABOOST
PRECISION	80%	62%	94%	92%
RECALL	89%	79%	79%	80%
F1-SCORE	84%	62%	84%	84%
ACCURACY	93,4%	74,4%	95,3%	95,2%

ELECTRASMALL

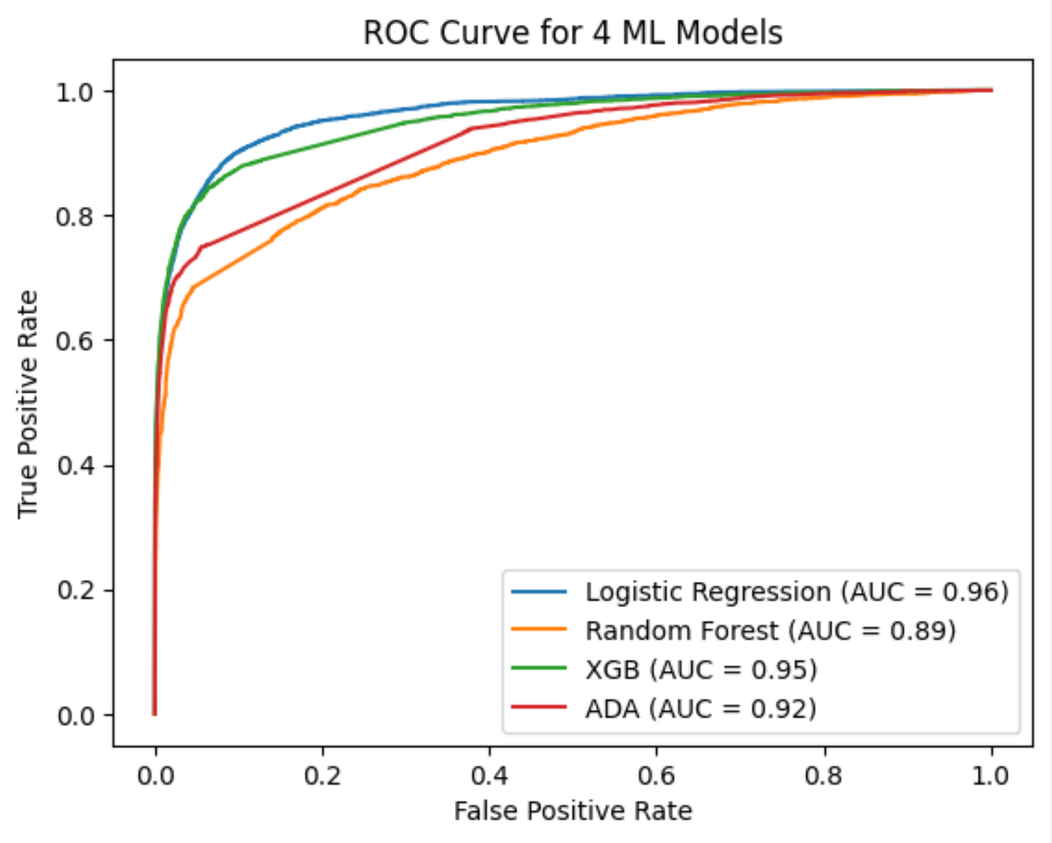
A **BERT-like model** pre-trained as a discriminator of a generative adversarial network (GAN):

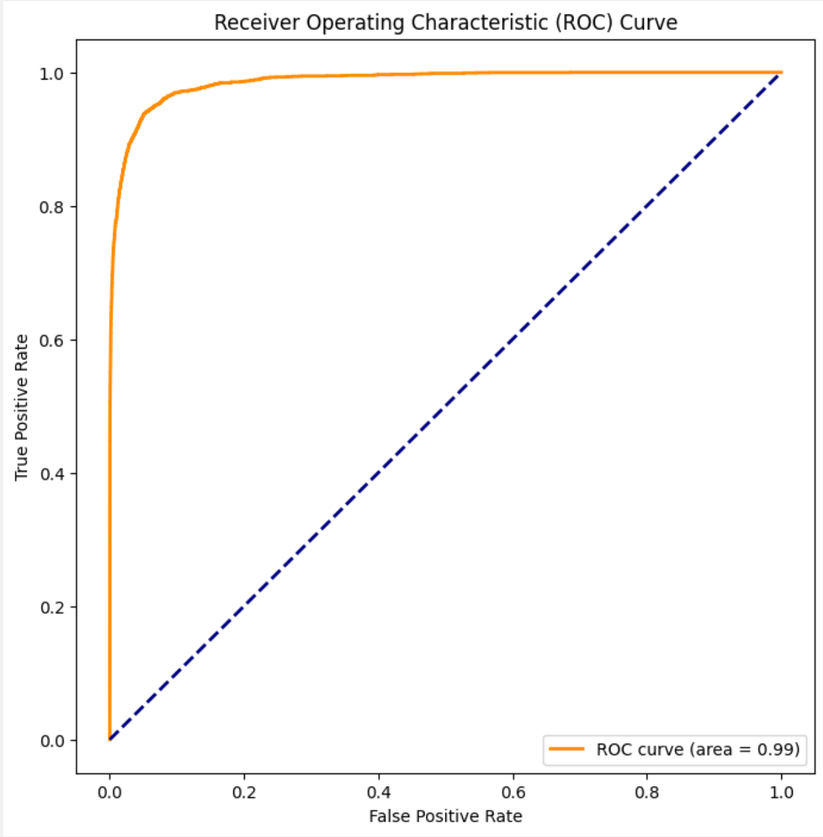
- text input: defines the input for the model. It takes in a string input of any shape.
- preprocessing layer: preprocess the text input. It points to the preprocessing module for the BERT model on TensorFlow Hub.
- encoder: load and utilize an encoder model. It point to the electra-Small model.
- dropout: applies a dropout regularization technique to the pooled output.
- classifier: fully connected dense layer with 1 unit (neurons) with sigmoid activation to produce an output ranging between 0 and 1.

ELECTRASMALL

- **Binary Cross-entropy as loss**: it evaluates the discrepancy between the probability distributions predicted by the model and the actual ones.
- **Binary Accuracy as metrics**: it measures the model's accuracy in predicting binary classes compared to the true labels. Essentially, it represents the fraction of samples correctly classified relative to the total.

RESULTS EVALUATION





TEST SET

Data points that the model has not encountered during training to avoid overfitting. It contains 153,164 comments.
No preprocessing phases is needed.

PREDICTION

TRAIN SET

TEXT CLUSTERING

It is an unsupervised learning approach, meaning that it does not require labeled data to train on.
Instead, the algorithm learns to identify patterns in the data and group documents based on those patterns.

WORD2VEC MODEL AND VECTORIZATION

word2vec groups documents together based on their overall meaning. Vectorization captures word relationships based on their co-occurrence in the corpus.

K-MEANS ALGORITHM

It aims to partition the data into a predefined number of clusters (k) by assigning data points to the nearest cluster centroid

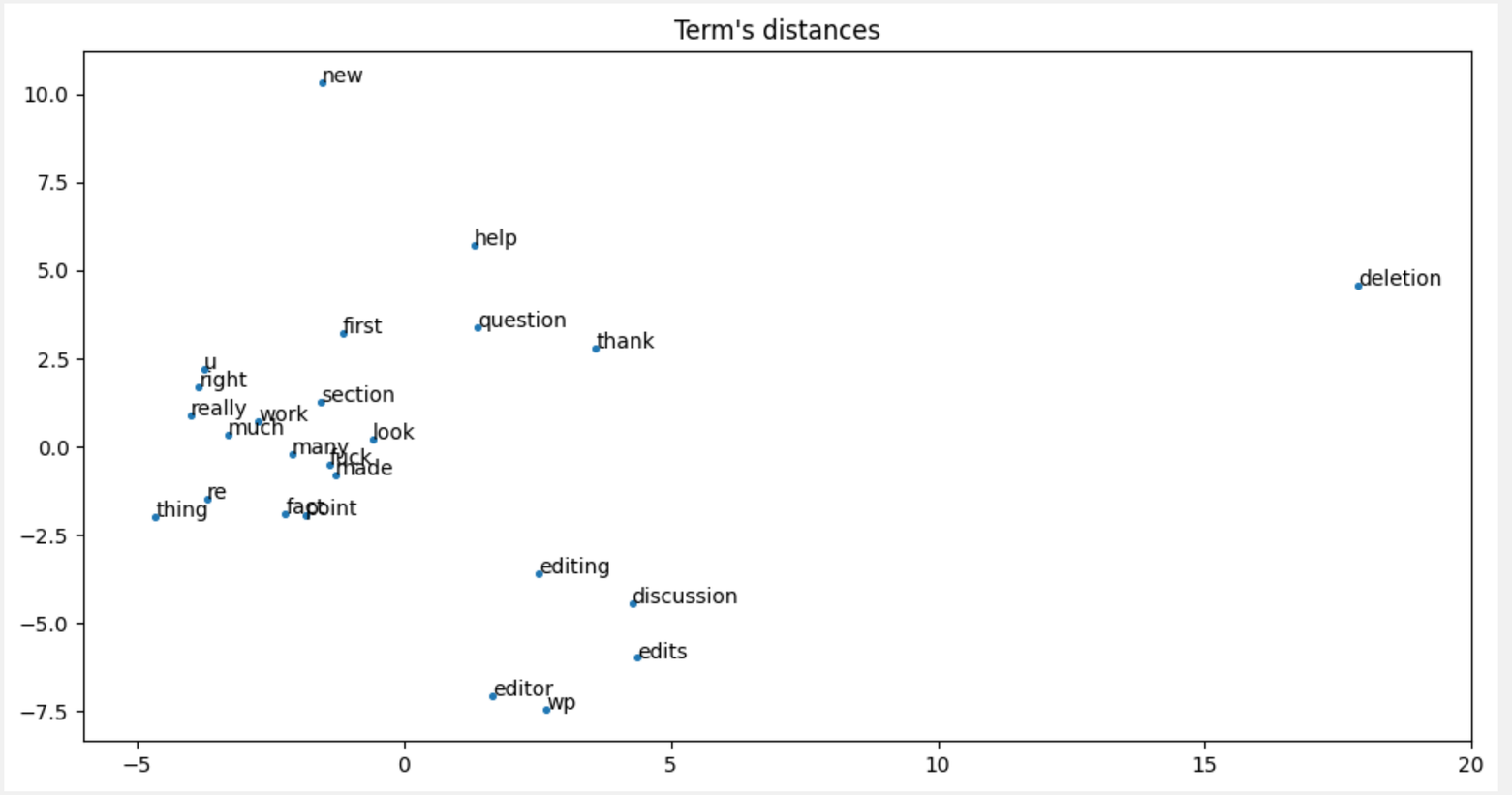
BOW, TF-IDF, SVD

The C-Means algorithm required its own preprocessing steps: Bag-of-Words, TF-IDF and SVD

FUZZY C-MEANS ALGORITHM (FCM)

C-means clustering is a generalization of k-means clustering that provide more flexibility.

WORD2VEC MODEL AND VECTORIZATION



However, the vectors produced by word2vec are not directly compatible with many machine learning algorithms, which typically require numerical data in a specific format. This is where vectorization techniques come into play. it captures word relationships based on their co-occurrence in the training corpus, effectively capturing contextual information. After the transformation, we obtain a list of vectors, one for each document and on length 300 (features).

K-MEANS ALGORITHM

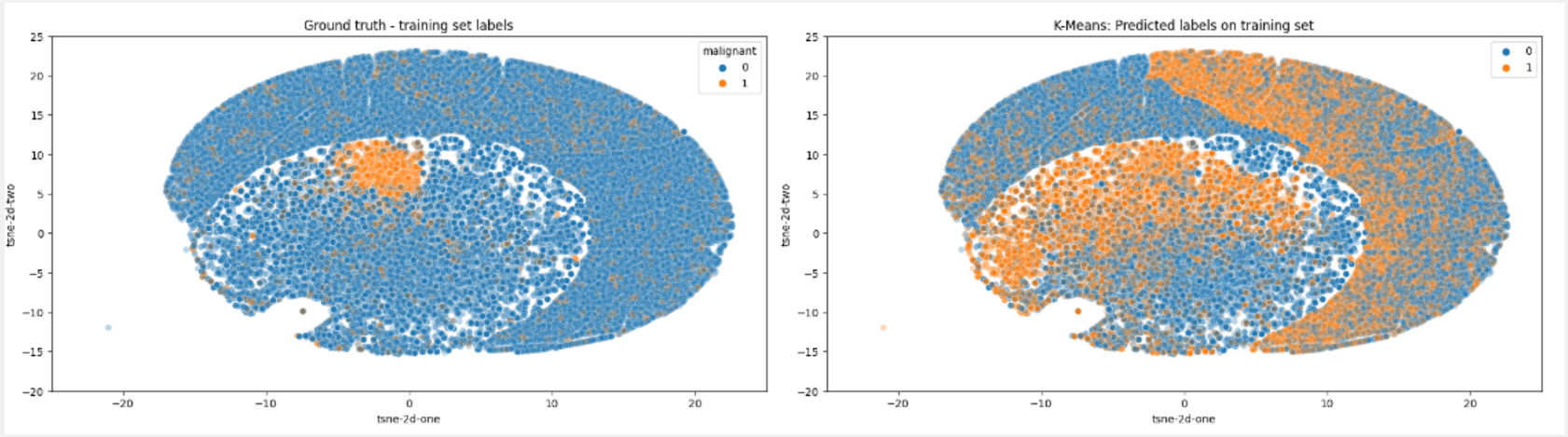
Hard clustering algorithms assign each data point to exactly one cluster, meaning that a data point can only belong to one cluster, and the membership of a data point in a cluster is either 1 or 0.

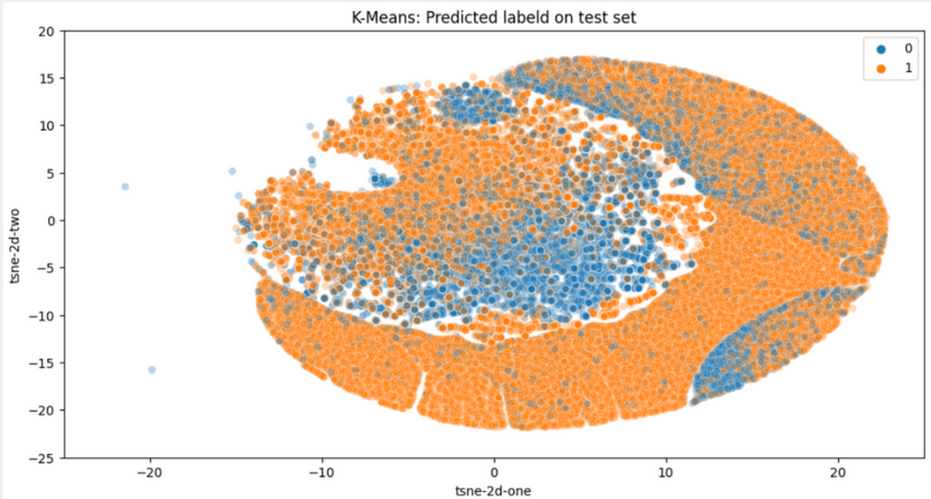
Silhouette coefficient value of 0.118:
the clustering results are not really strong. This means that the clusters are not well-separated and that there is a high degree of overlap between them.

Cluster	Most Representative Terms
0	brag, fatass, confront, ure, mokele
1	article, anyway, necessary, particular, otherwise

Cluster	Size
0	54728
1	98436

K-MEANS ALGORITHM



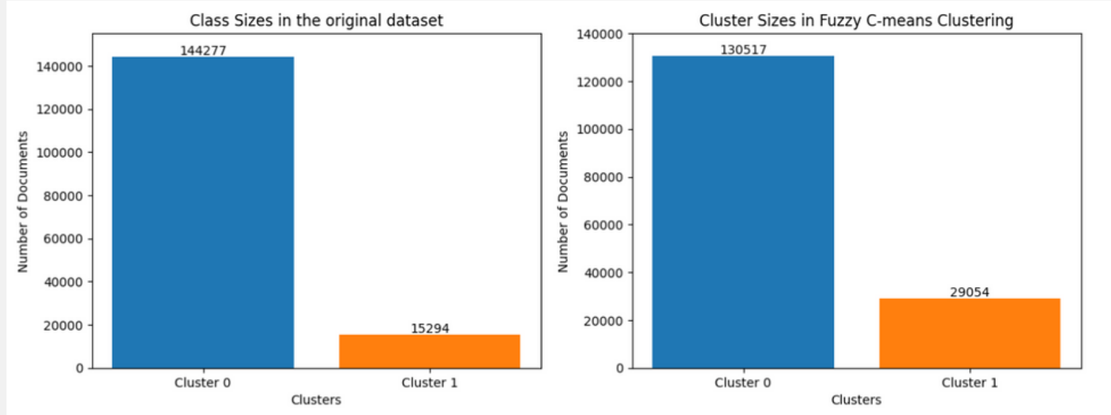


CERVINI STELLA, SABINO GIUSEPPE

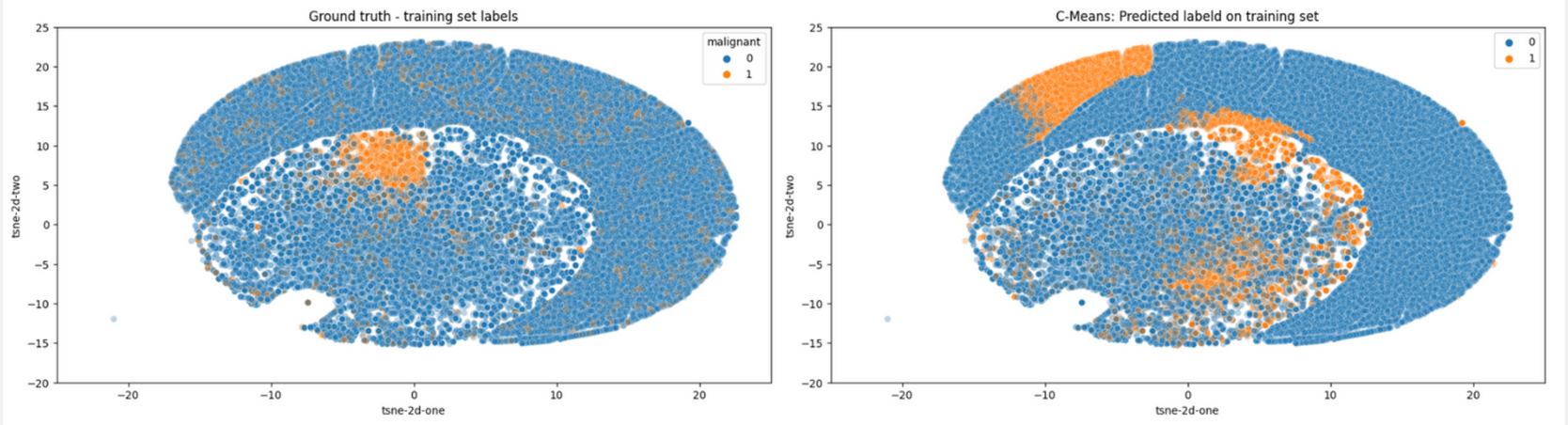
FUZZY C-MEANS ALGORITHM

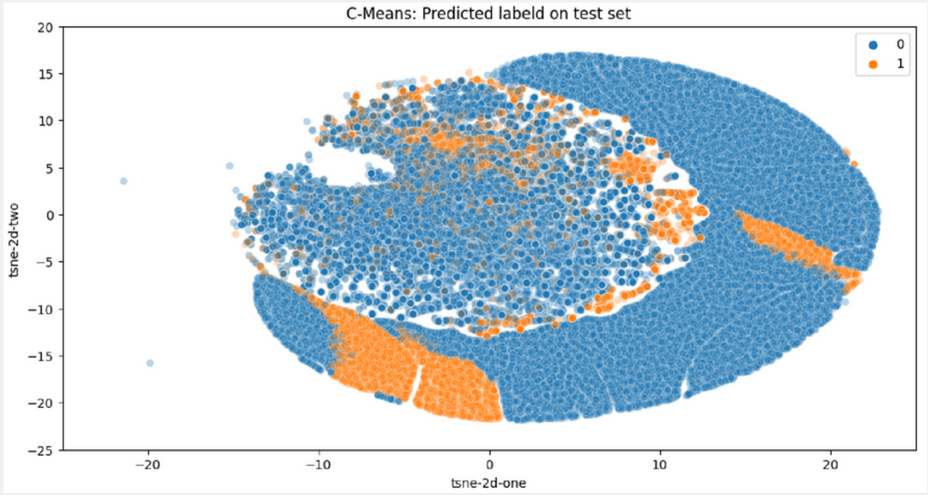
Soft clustering algorithms assign each data point to a distribution over all clusters. Thus, a data point can belong to multiple clusters, and the membership of a data point in a cluster can be any value between 0 and 1.

FPC is approximately 0.82:
it suggests that the fuzzy clustering result is favorable. A higher FPC suggests a well-defined and separate groups. On the contrary, a lower FPC hints that the boundaries between clusters are more diffuse, and data points exhibit a higher degree of ambiguity.



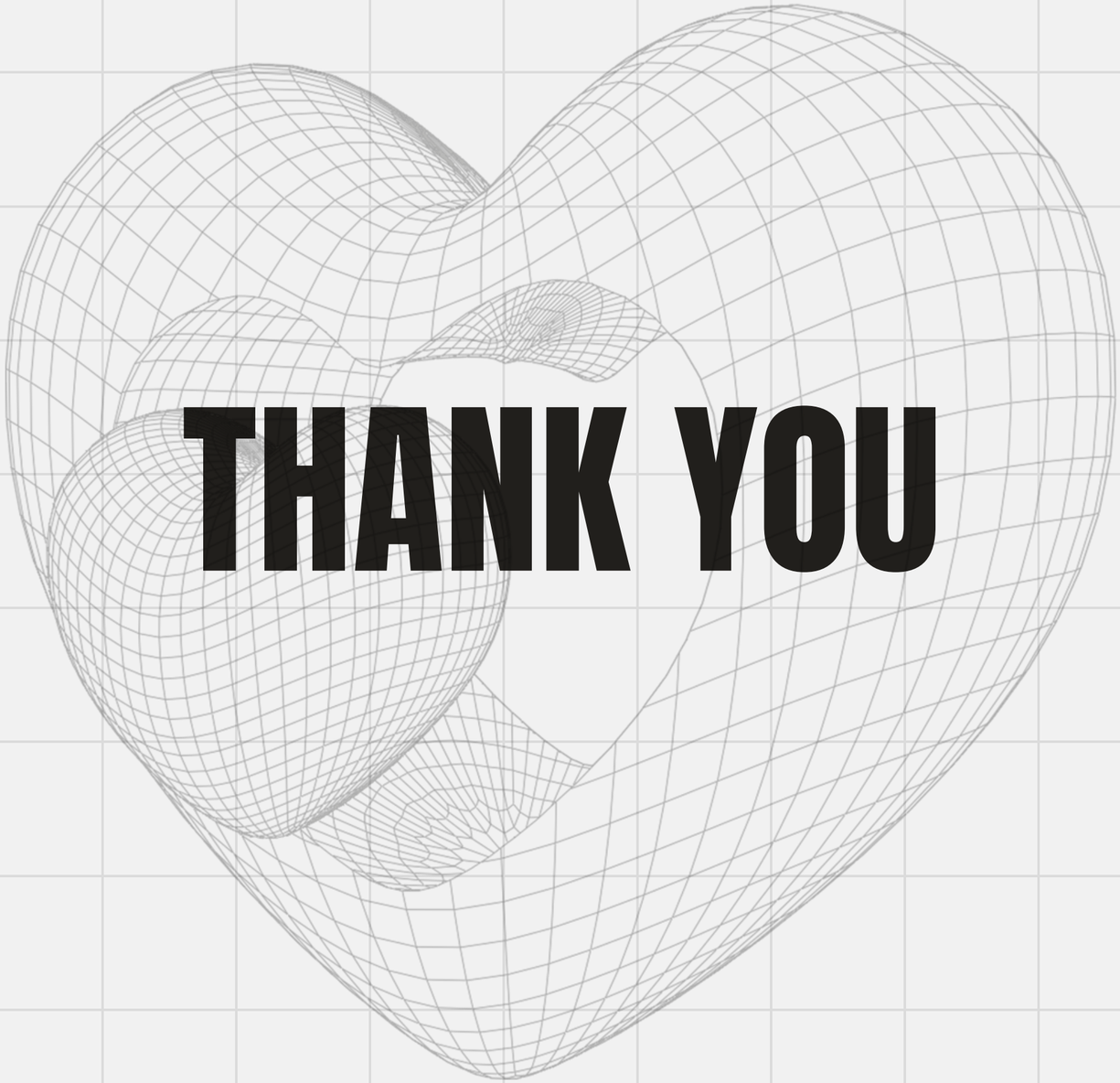
FUZZY C-MEANS ALGORITHM





CONLUSIONS

Malignant comment identification using classification and clustering techniques demonstrates the effectiveness of machine learning in identifying and characterizing harmful online content. These conclusions highlight the potential of machine learning in addressing online toxicity and safeguarding online communities. As these techniques continue to evolve, their role in combatting online abuse and promoting safe and inclusive online spaces is expected to grow significantly.



THANK YOU