# Decision Trees

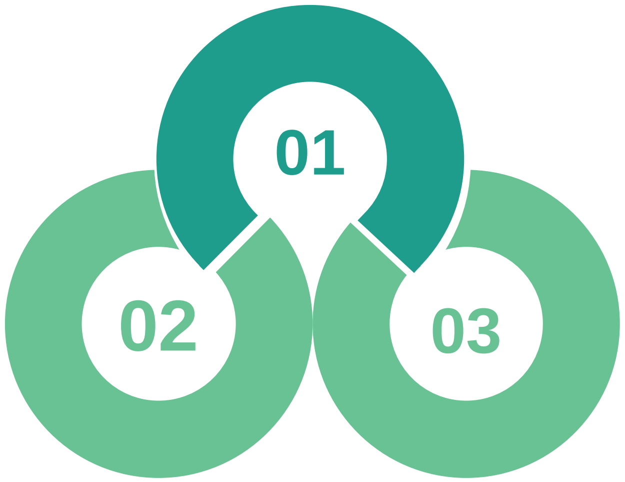
Docent: Dr. Stefan Langer
Student: Zonghao Yang
Seminar: Seminar Klassifikation und Clustering
Date:  29th January 2024

# CONTENTS

# PART 1

Introduction

# ADD YOUR TITLE HERE




1

## Early Developments

The conceptual foundation of decision trees dates back to the 1950s and 1960s.

The modern form of decision trees, as used in machine learning, started to take shape in the 1980s.

2

## Evolution and Adoption

Over the years, decision trees have evolved greatly, with advancements in algorithms to handle overfitting, better ways to split nodes, and integration with other machine learning techniques like ensemble methods.

Random Forests is an ensemble method that constructs a multitude of decision trees at training time.

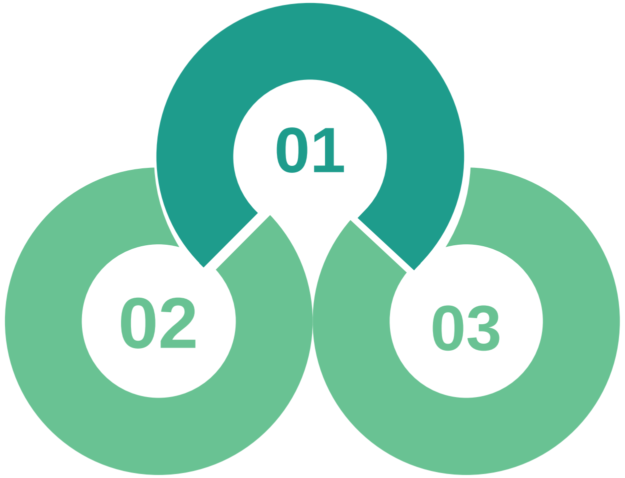
3

## Modern Usage

Decision trees are widely used in various domains for classification tasks due to their simplicity and effectiveness. They are the basis for many advanced machine learning models.

The history of decision trees shows the evolution of a simple and powerful idea into a foundation of modern machine learning and data analysis techniques.

# Milestones




1

## ID3(Iterative Dichotomiser 3)

The development of the ID3 algorithm in the 1980s was an important milestone in the history of decision trees.

ID3 was one of the first algorithms to employ a top-down, greedy search through the given sets to construct a decision tree.

2

## C4.5

introduced in 1993,  ID3 was further improved upon with the C4.5 algorithm.

C4.5 made several enhancements over ID3, such as the ability to handle both continuous and discrete attributes, deal with missing data, and prune trees after construction.

3

## CART (Classification and Regression Trees)

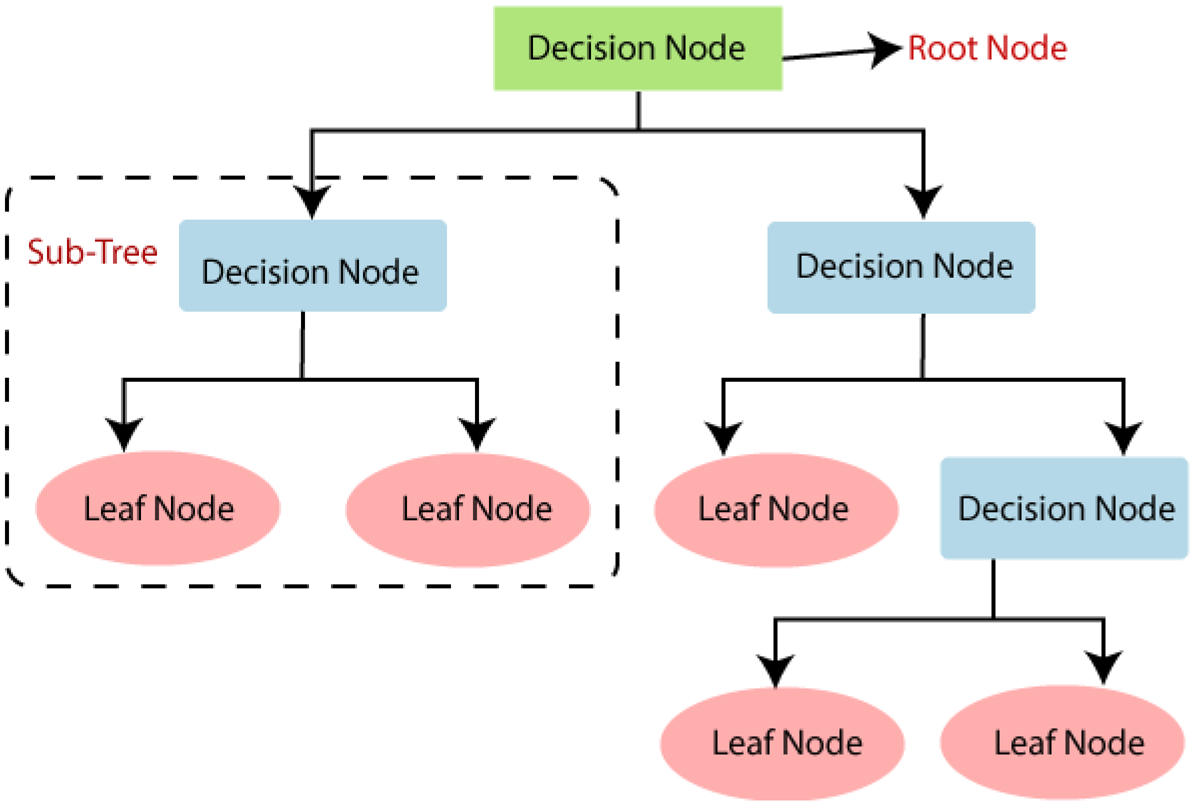
CART was developed in 1984, almost at the same time as ID3.

CART introduced the concept of constructing binary trees, focusing on selecting the feature and threshold that lead to the most significant reduction in Gini impurity at each node, while ID3 and C4.5 use entropy and information gain as criteria.

# PART 2

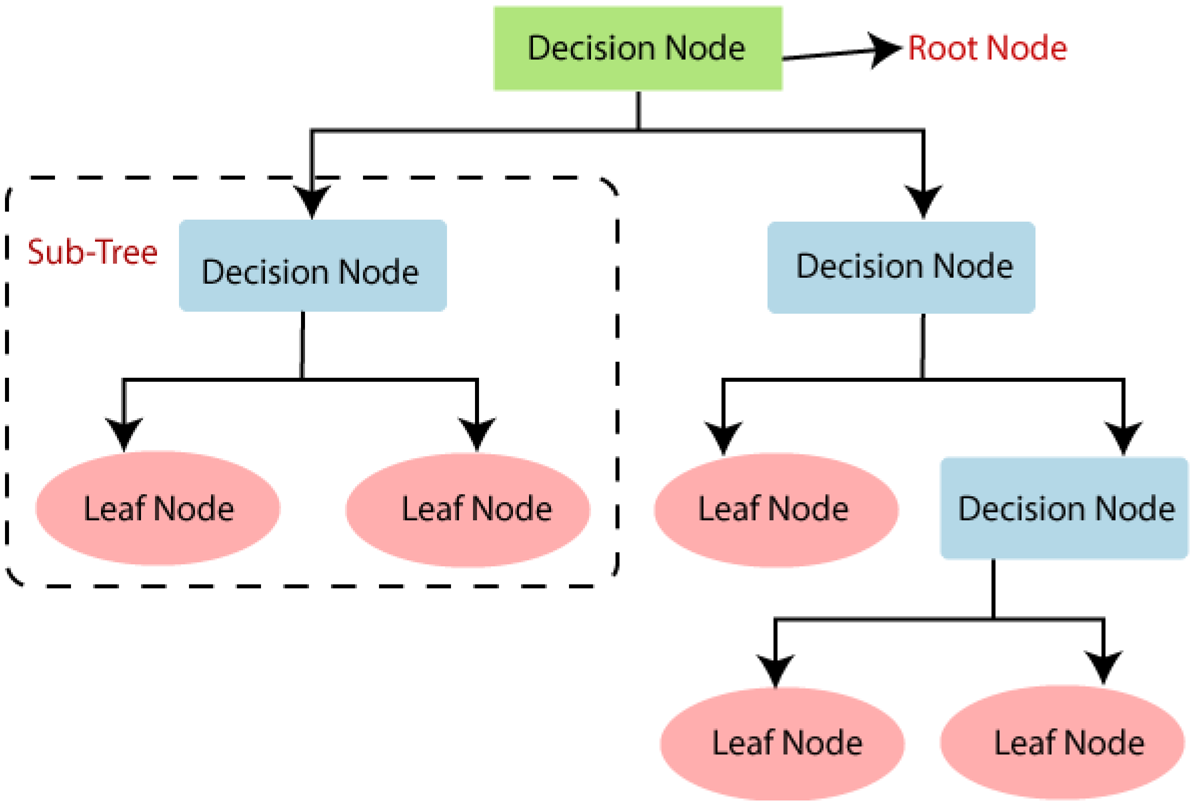
Concepts of Decision trees

# Structure of Decision Tree




- **Root Node: Represents the entire data and splits into two or more homogeneous sets.**
- **Splitting: The process of dividing a node into two (CART) or more sub-nodes(ID3).**
- **Decision Node: A sub-node that can split into sub-nodes further.**
- **Leaf Node: Nodes that do not split; they represent the final output or decision.**
- **Branch: A subsection of the entire tree.**

# Structure of Decision Tree




**How does the Decision Tree algorithm Work?**

**Step 1: Begin the tree with the root node, which contains the complete dataset.**

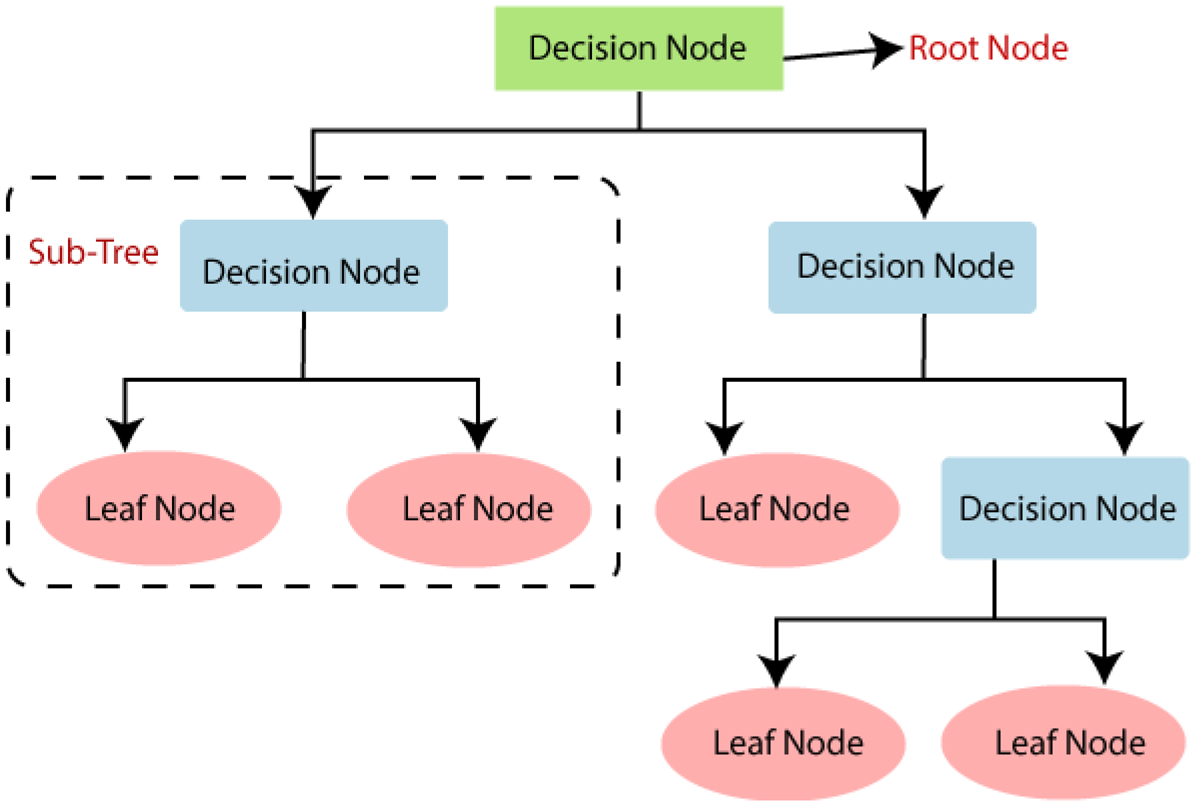
**Step 2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).**

**Step 3: Divide the root nide into subsets that contains possible values for the best attributes.**

**Step 4: Generate the decision tree node, which contains the best attribute.**

**Step 5: Recursively make new decision trees using the subsets of the dataset created in step -3.**
**Continue this process until a stage is reached where you cannot further classify the nodes.**

# Working Mechanism




**Feature Selection:** The decision of which feature to split at each node is very important.

This is determined using statistical measures like Gini impurity, Entropy and Information Gain.

**Splitting Criteria:**

Gini index is a measure of impurity used while creating a decision tree in the CART algorithm to create binary splits.

An attribute with the low Gini index should be preferred as compared to the high Gini index.

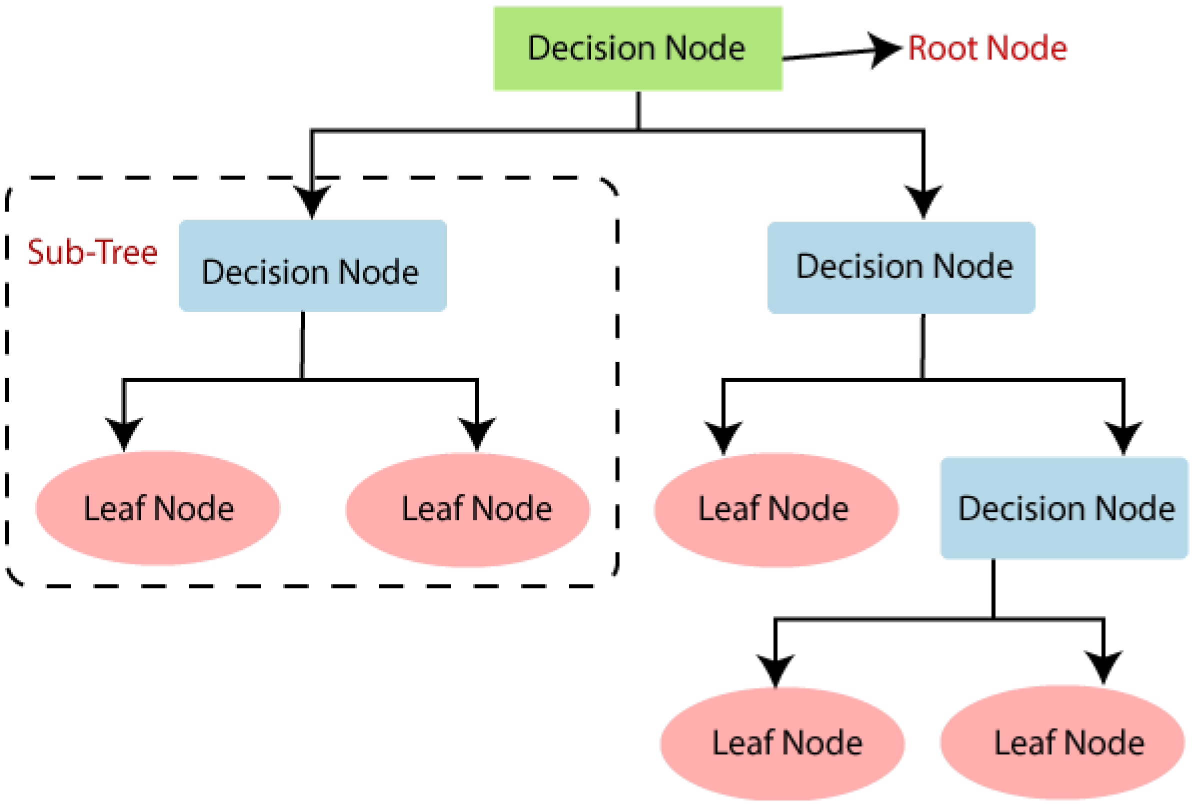
Information Gain is the reduction in entropy obtained by dividing the dataset according to a particular feature.

It is used to determine which feature to split on at each step in building a decision tree.

The model compares every possible split and pick the one with the highest information gain.

Higher information gain suggests the more useful feature for creating distinct groups.

# P r u n i n g




**Pruning is the process of removing parts of the decision tree that are unnecessary or less important for making predictions.**

**This process reduces the tree's complexity without reducing its accuracy too much.**

**Why is Pruning Important?**

**Prevents Overfitting: Larger trees can overfit the training data and capture noise.**

**Improves Model Simplicity: A smaller tree is easier to understand and interpret.**

**Pruning can be achieved by setting limits on the maximum depth of the tree, the minimum number of samples required to split a node and the minimum number of samples required in a leaf.**

# PART 3

Advantages and Disadvantages

# Advantages

1 **Easy to Understand and Interpret:**

**Decision trees can be visualized and understood easily.**

2 **Good at handling Non-linear Relationships:**

**Decision trees perform well in handling non-linear relationships in data, as they segment the space into smaller sub-spaces based on the features.**

3 **Flexibility:**

**Decision trees can be used for both classification and regression tasks.**

**Their performance can be enhanced through ensemble methods like Random Forests.**

# Disadvantages

1 **Overfitting:**
Trees can create over-complex models that do not generalize well to new data. Pruning methods can help.

2 **Instability:**
Small variations in data can result in a completely different tree.

3 **Biased Trees with Imbalanced Datasets:**
Decision trees can be biased toward the dominant class.

4 **Difficulty in Capturing Complex Relationships:**
While good for simple tasks, decision trees can struggle with tasks requiring the modeling of more complex relationships.

5 **Not fit for Large Datasets:**
As the size of the dataset increases, the complexity of decision trees can grow, making them impractical for very large datasets.

# PART 4

Results

# p a r a m e t e r s

1 **max_depth:**

Controls the maximum depth of the tree. Limiting the depth can help prevent overfitting by reducing the complexity of the model.

2 **min_samples_split:**

The number of samples required to split an internal node. Adjusting this parameter can help control the number of splits and thus the complexity of the tree.

3 **min_samples_leaf:**

The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least the given amount of training samples in each of the left and right branches.

4 **max_features:**

Determines the number of features to consider when looking for the best split.

5 **criterion:**

Gini impurity is the default criterion, "entropy" is also an option.

6 **splitter:**

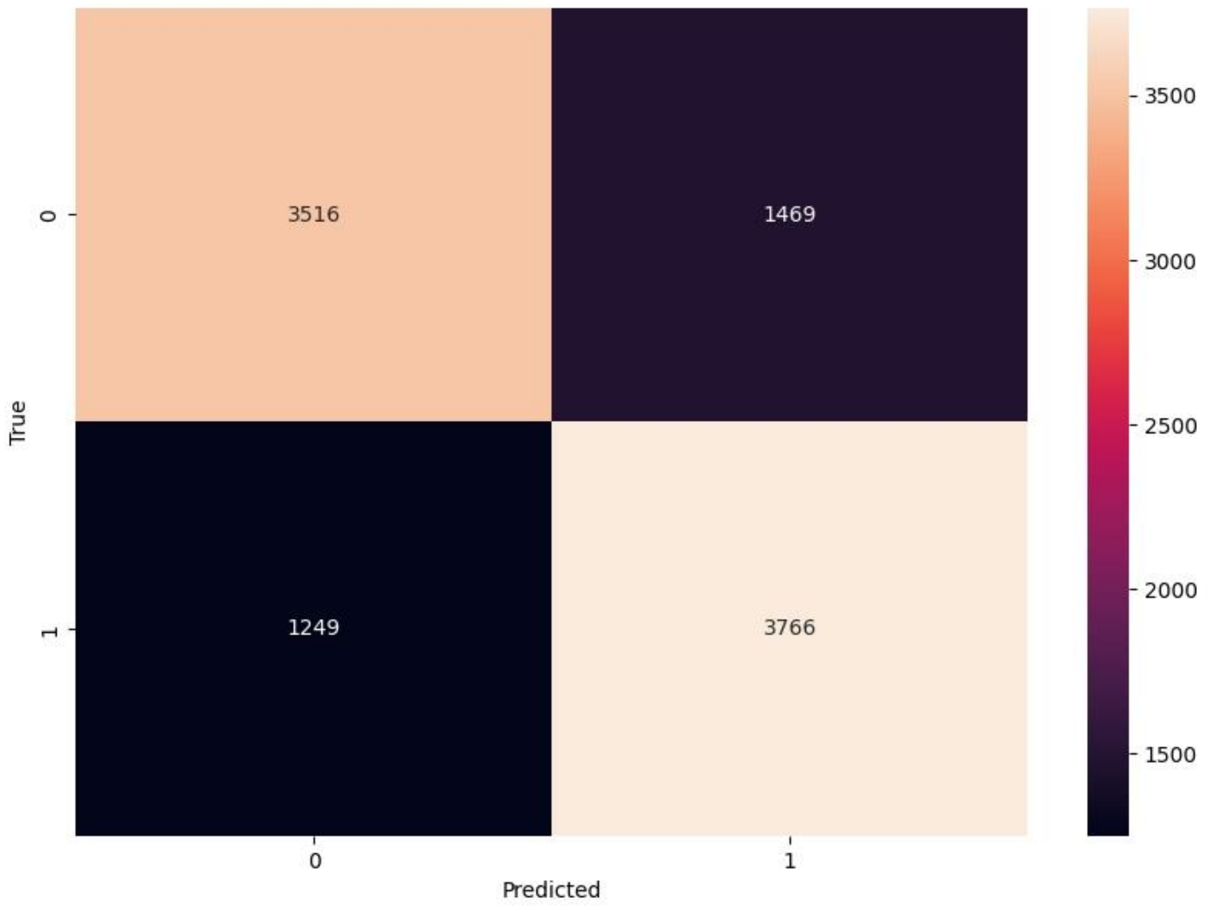
The strategy for choosing the split at each node.

- min_samples_split=20, max_features=None, min_samples_leaf=5, max_depth=20
- tf- idf(unigram)

**Accuracy: 0.73**

**Macro Average:0.73 Weighted Average:0.73**

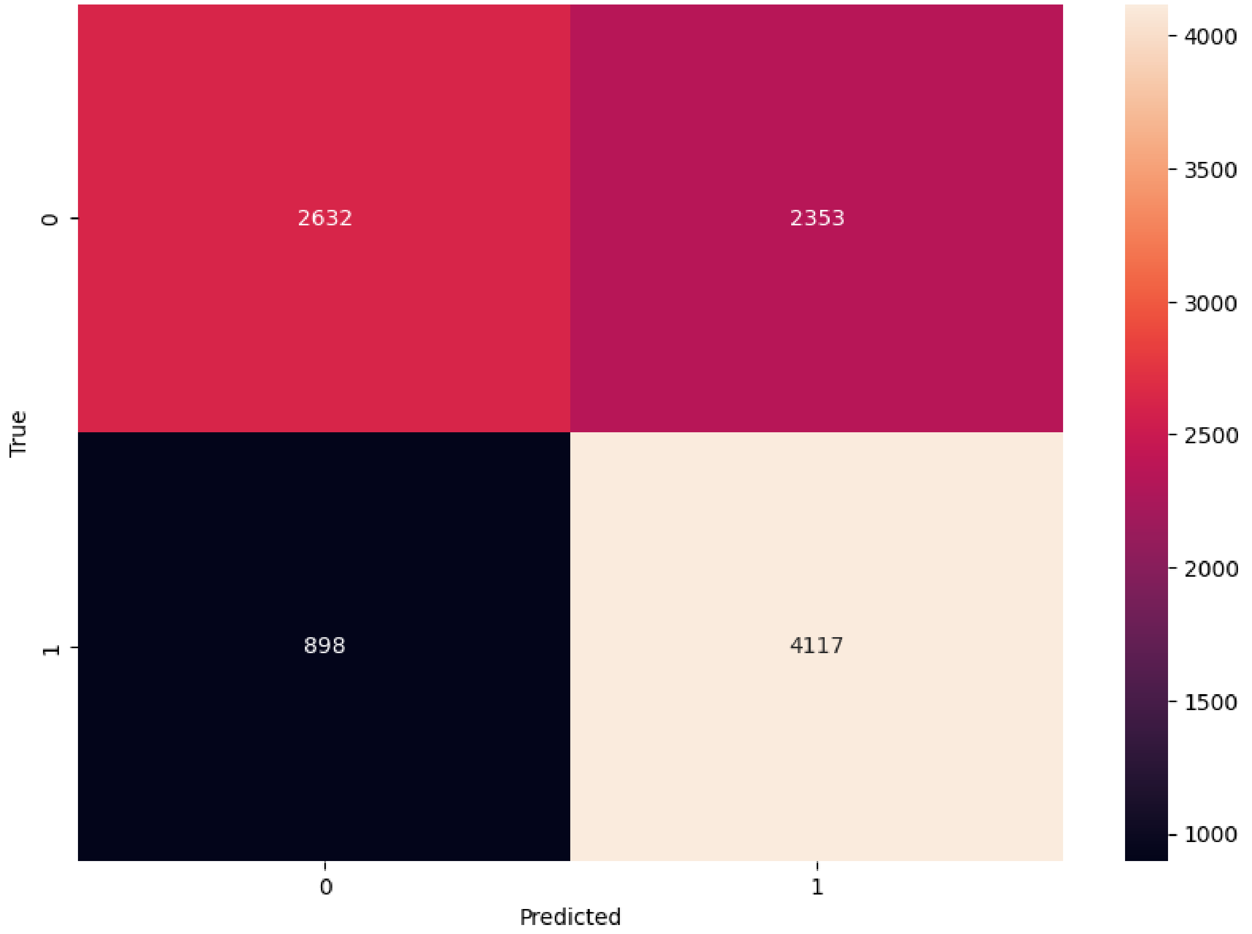
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negative | 0.74 | 0.71 | 0.72 | 4985 |
| positive | 0.72 | 0.75 | 0.73 | 5015 |
| accuracy | | | 0.73 | 10000 |
| macro avg | 0.73 | 0.73 | 0.73 | 10000 |
| weighted avg | 0.73 | 0.73 | 0.73 | 10000 |



- min_samples_split=20, max_features=None, min_samples_leaf=5, max_depth=20
- tf- idf(bigram)

**Accuracy: 0.67** **Macro Average:0.67 Weighted Average:0.67**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negative | 0.75 | 0.53 | 0.62 | 4985 |
| positive | 0.64 | 0.82 | 0.72 | 5015 |
| accuracy | | | 0.67 | 10000 |
| macro avg | 0.69 | 0.67 | 0.67 | 10000 |
| weighted avg | 0.69 | 0.67 | 0.67 | 10000 |

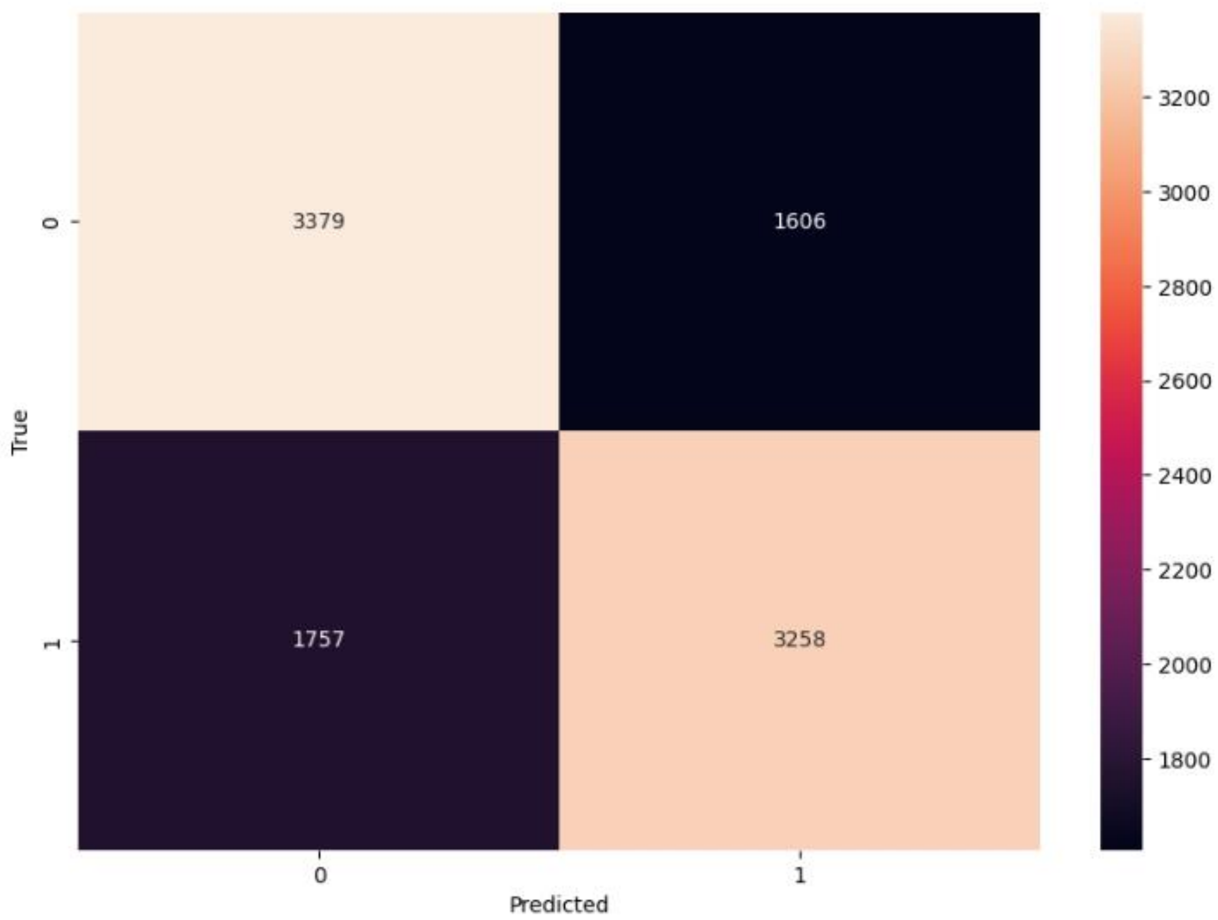


- min_samples_split=20, max_features=None, min_samples_leaf=5, max_depth=20

- Word2Vec

- Accuracy: 0.66
- F1-Score: Positive 0.66; Negative 0.65

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| negative | 0.66 | 0.68 | 0.67 | 4985 |
| positive | 0.67 | 0.65 | 0.66 | 5015 |
| accuracy | | | 0.66 | 10000 |
| macro avg | 0.66 | 0.66 | 0.66 | 10000 |
| weighted avg | 0.66 | 0.66 | 0.66 | 10000 |



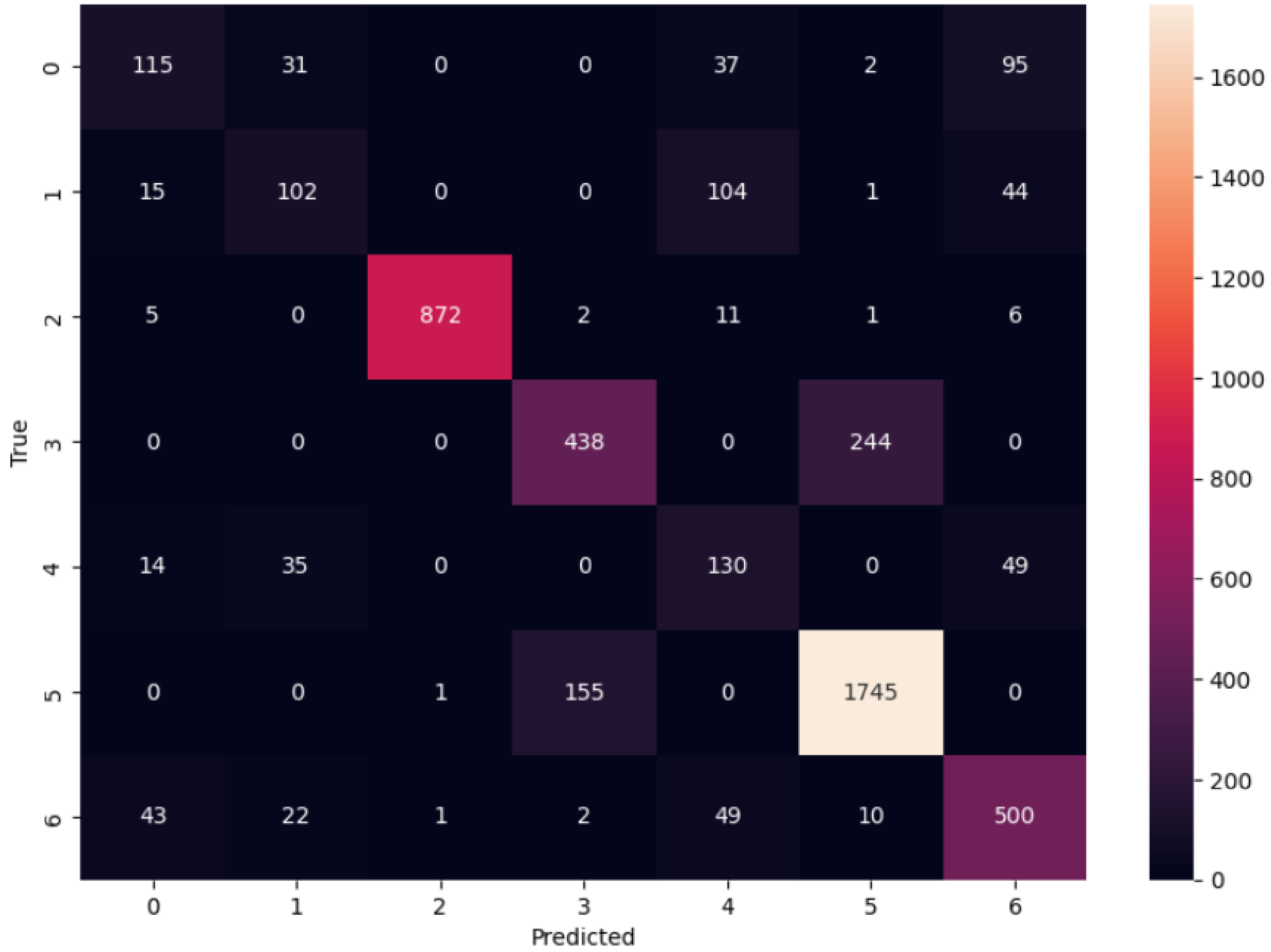
# Letter(author)

- min_samples_split=20, max_features=None, min_samples_leaf=5, max_depth=30
- tf-idf (unigram)

**Accuracy: 0.80**

**Macro Average:0.67 Weighted Average:0.80**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Franz Kafka | 0.60 | 0.41 | 0.49 | 280 |
| Friedrich Schiller | 0.54 | 0.38 | 0.45 | 266 |
| Henrik Ibsen | 1.00 | 0.97 | 0.98 | 897 |
| James Joyce | 0.73 | 0.64 | 0.68 | 682 |
| Johann Wolfgang von Goethe | 0.39 | 0.57 | 0.47 | 228 |
| Virginia Woolf | 0.87 | 0.92 | 0.89 | 1901 |
| Wilhelm Busch | 0.72 | 0.80 | 0.76 | 627 |
| accuracy | | | 0.80 | 4881 |
| macro avg | 0.69 | 0.67 | 0.67 | 4881 |
| weighted avg | 0.80 | 0.80 | 0.80 | 4881 |



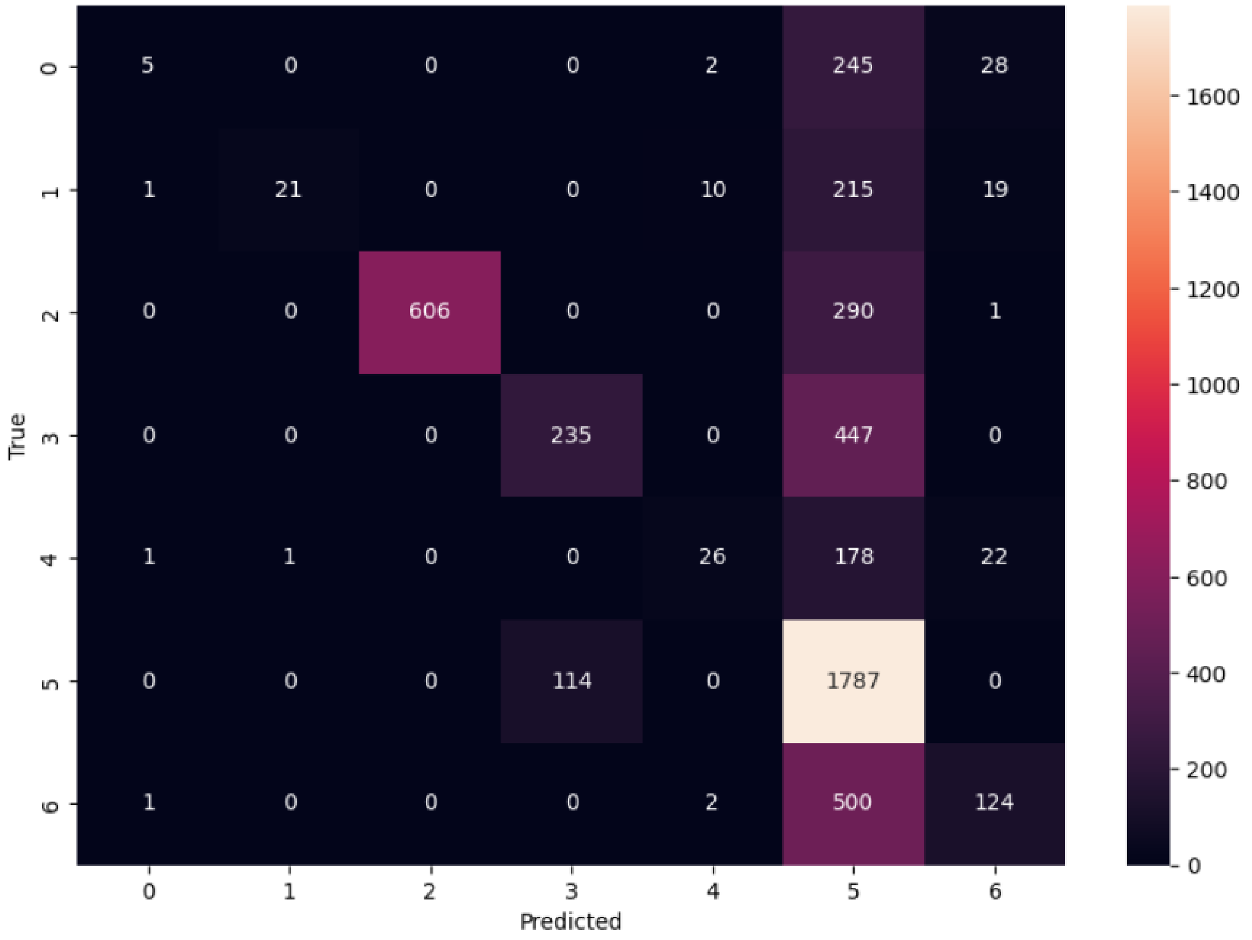
- min_samples_split=20, max_features=None, min_samples_leaf=5, max_depth=30

- tf-idf (bigram)

- **Accuracy: 0.80**

- **Macro Average:0.67 Weighted Average:0.80**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Franz Kafka | 0.62 | 0.02 | 0.03 | 280 |
| Friedrich Schiller | 0.95 | 0.08 | 0.15 | 266 |
| Henrik Ibsen | 1.00 | 0.68 | 0.81 | 897 |
| James Joyce | 0.67 | 0.34 | 0.46 | 682 |
| ohann Wolfgang von Goethe | 0.65 | 0.11 | 0.19 | 228 |
| Virginia Woolf | 0.49 | 0.94 | 0.64 | 1901 |
| Wilhelm Busch | 0.64 | 0.20 | 0.30 | 627 |
| accuracy | | | 0.57 | 4881 |
| macro avg | 0.72 | 0.34 | 0.37 | 4881 |
| weighted avg | 0.67 | 0.57 | 0.52 | 4881 |



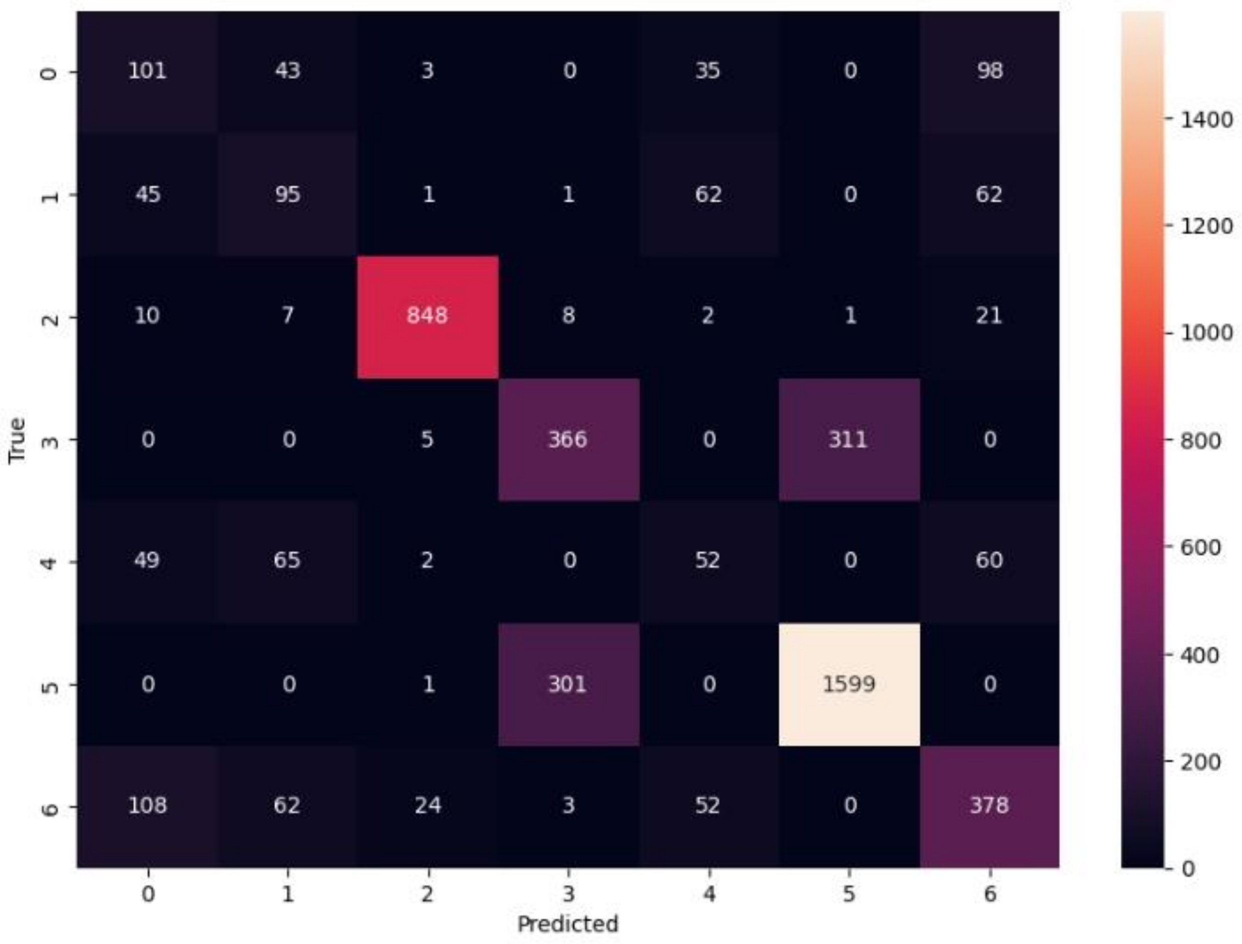
## min_samples_split=20, max_features=None, min_samples_leaf=5, max_depth=30

### Word2Vec

**Accuracy: 0.70**

**Macro Average:0.55 Weighted Average:0.70**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Franz Kafka | 0.32 | 0.36 | 0.34 | 280 |
| Friedrich Schiller | 0.35 | 0.36 | 0.35 | 266 |
| Henrik Ibsen | 0.96 | 0.95 | 0.95 | 897 |
| James Joyce | 0.54 | 0.54 | 0.54 | 682 |
| ohann Wolfgang von Goethe | 0.26 | 0.23 | 0.24 | 228 |
| Virginia Woolf | 0.84 | 0.84 | 0.84 | 1901 |
| Wilhelm Busch | 0.61 | 0.60 | 0.61 | 627 |
| accuracy | | | 0.70 | 4881 |
| macro avg | 0.55 | 0.55 | 0.55 | 4881 |
| weighted avg | 0.71 | 0.70 | 0.70 | 4881 |

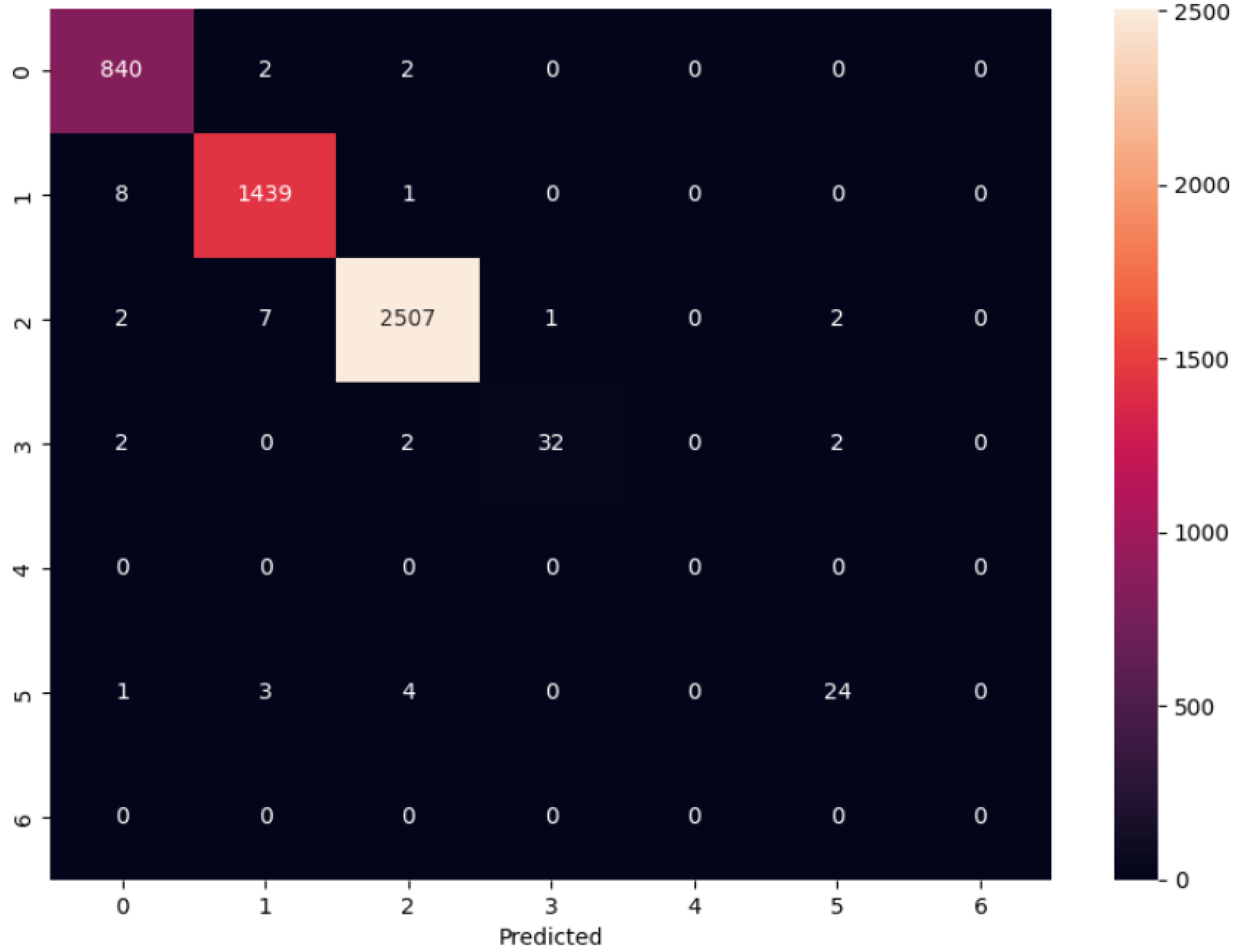


- min_samples_split=20, max_features=None, min_samples_leaf=5, max_depth=20

- tf-idf (unigram)

- **Accuracy: 0.99**
- **Macro Average:0.94 Weighted Average:0.99**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| da | 0.98 | 1.00 | 0.99 | 844 |
| de | 0.99 | 0.99 | 0.99 | 1448 |
| en | 1.00 | 1.00 | 1.00 | 2519 |
| fr | 0.97 | 0.84 | 0.90 | 38 |
| it | 0.86 | 0.75 | 0.80 | 32 |
| accuracy | | | 0.99 | 4881 |
| macro avg | 0.96 | 0.92 | 0.94 | 4881 |
| eighted avg | 0.99 | 0.99 | 0.99 | 4881 |



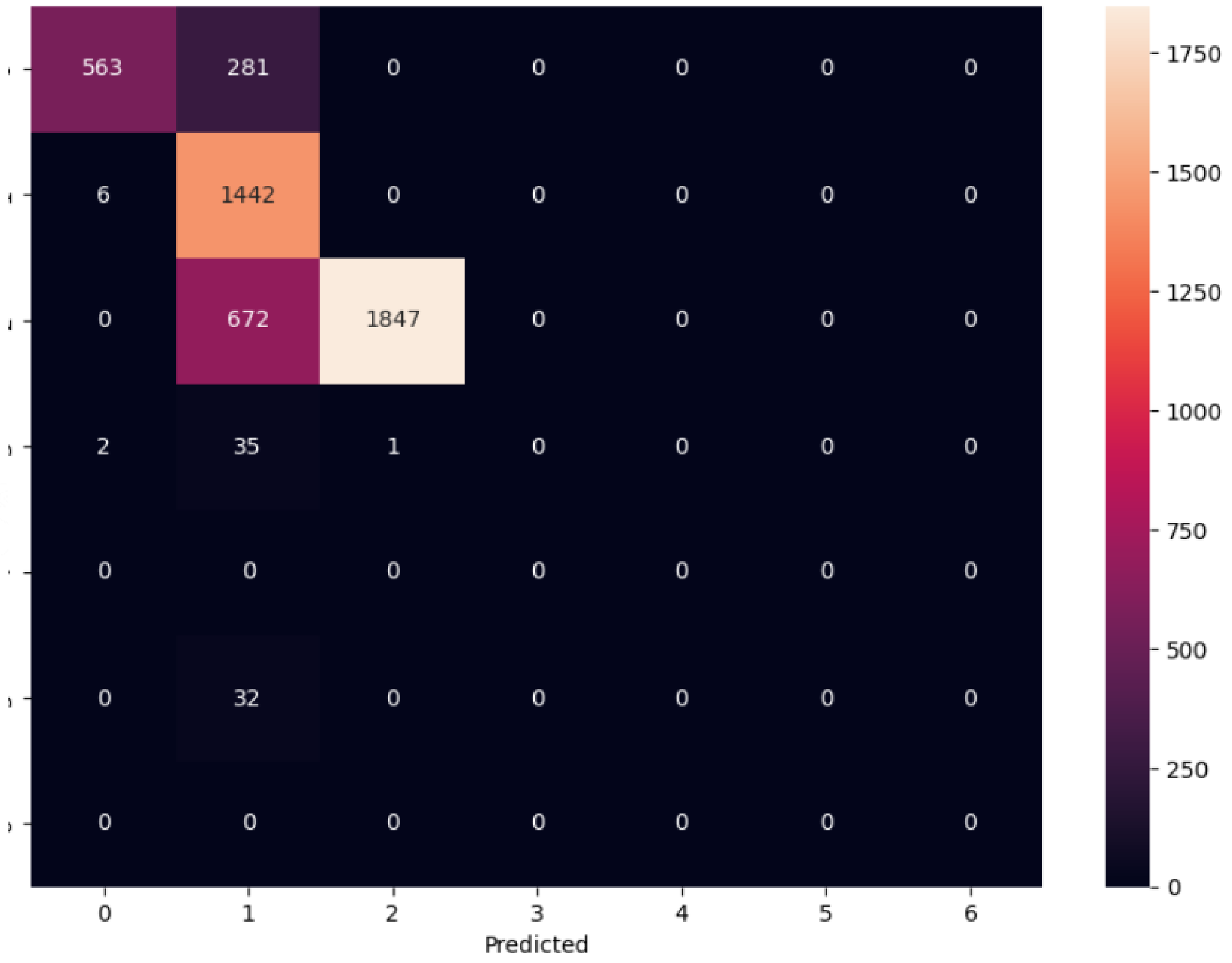
# Letter

min_samples_split=10, max_features=None, min_samples_leaf=5, max_depth=30

tf-idf (bigram)

**Accuracy: 0.79** **Macro Average:0.48 Weighted Average:0.79**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| da | 0.99 | 0.67 | 0.80 | 844 |
| de | 0.59 | 1.00 | 0.74 | 1448 |
| en | 1.00 | 0.73 | 0.85 | 2519 |
| fr | 0.00 | 0.00 | 0.00 | 38 |
| it | 0.00 | 0.00 | 0.00 | 32 |
| accuracy | | | 0.79 | 4881 |
| macro avg | 0.51 | 0.48 | 0.48 | 4881 |
| eighted avg | 0.86 | 0.79 | 0.79 | 4881 |

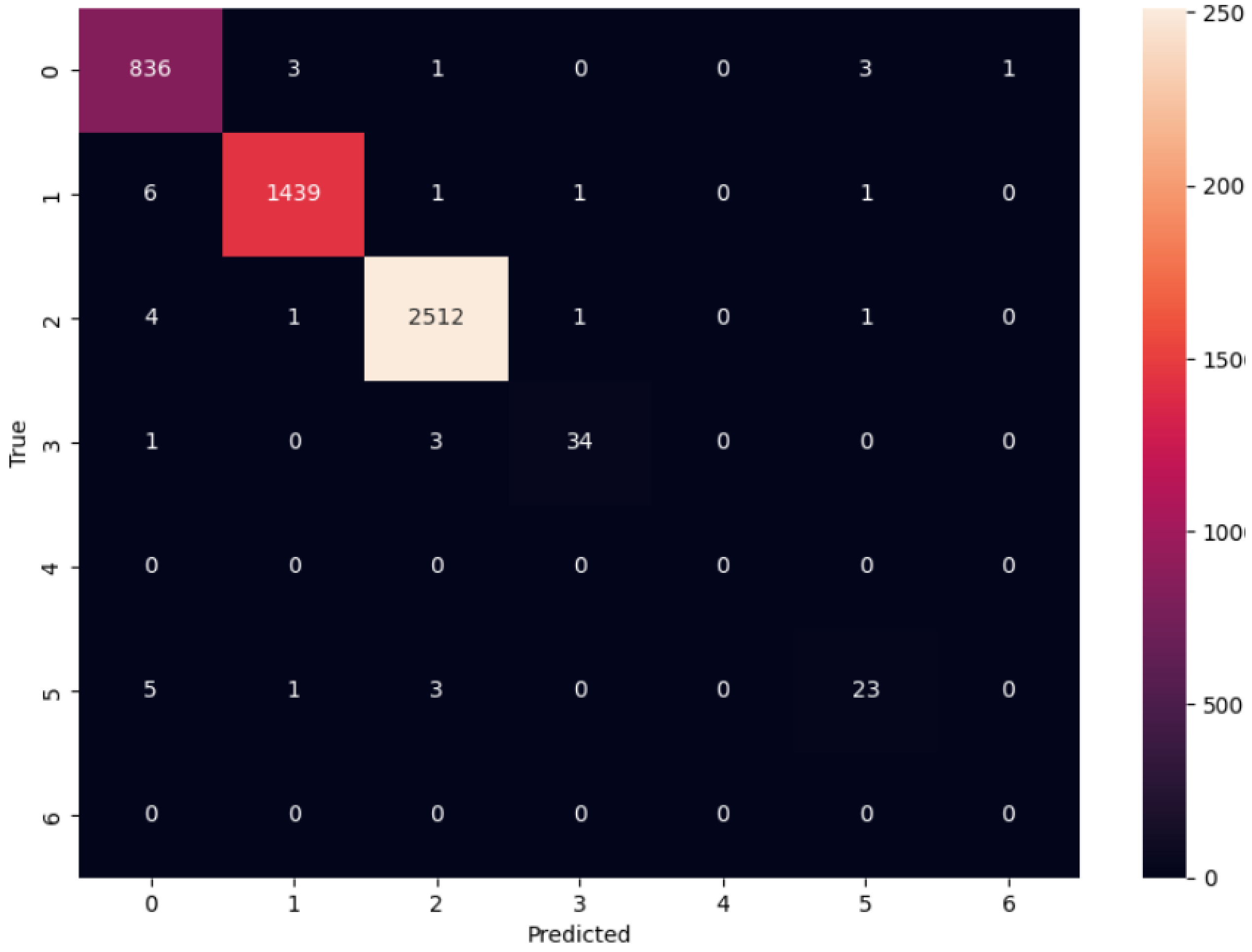


- min_samples_split=10, max_features="sqrt"

- Word2Vec

- **Accuracy: 0.99**
- **Macro Average:0.78 Weighted Average:0.99**

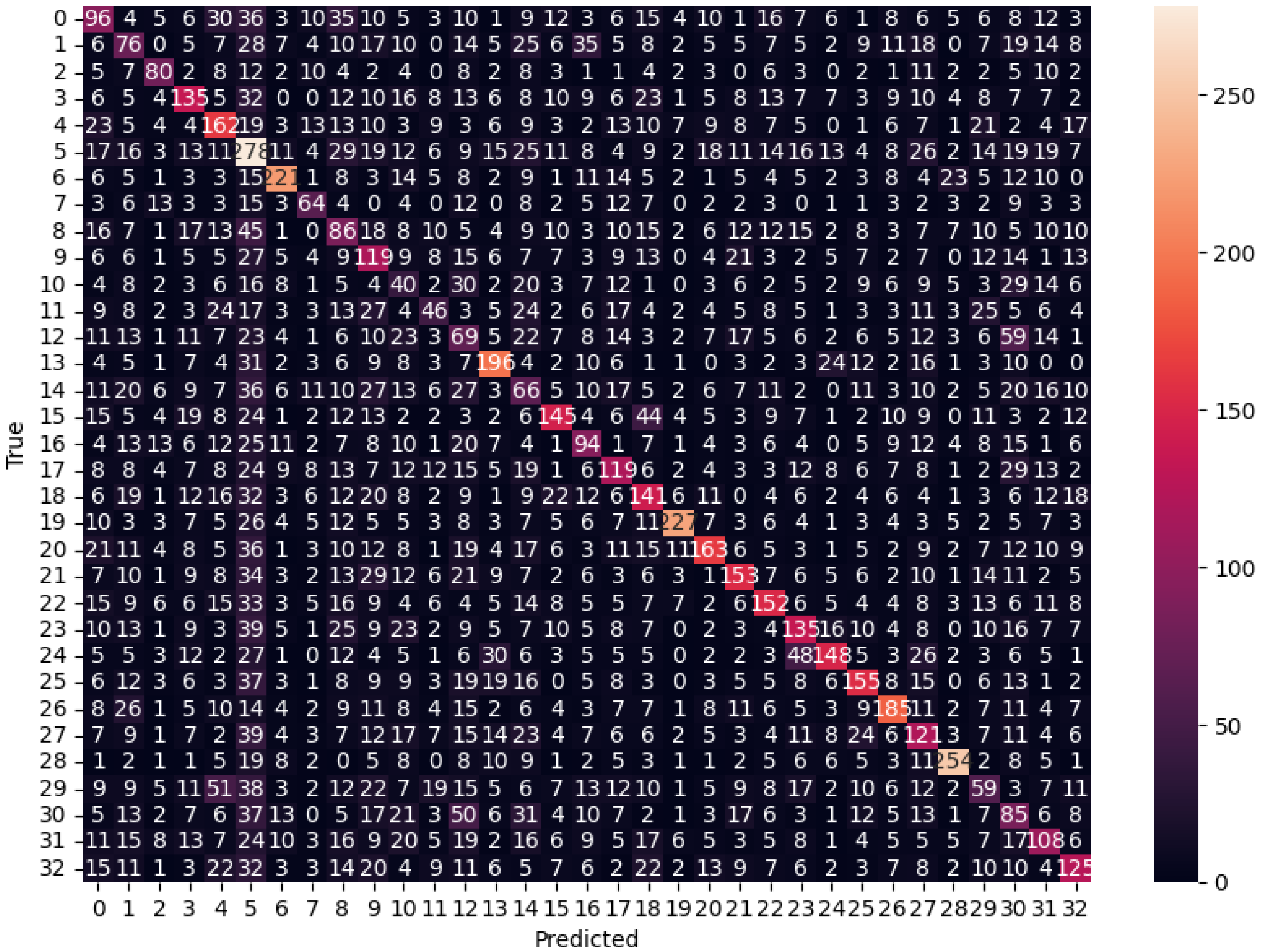
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| da | 0.98 | 0.99 | 0.99 | 844 |
| de | 1.00 | 0.99 | 1.00 | 1448 |
| en | 1.00 | 1.00 | 1.00 | 2519 |
| fr | 0.94 | 0.89 | 0.92 | 38 |
| it | 0.82 | 0.72 | 0.77 | 32 |
| sv | 0.00 | 0.00 | 0.00 | 0 |
| accuracy | | | 0.99 | 4881 |
| macro avg | 0.79 | 0.77 | 0.78 | 4881 |
| weighted avg | 0.99 | 0.99 | 0.99 | 4881 |



## min_samples_split=50, max_features=None, min_samples_leaf=5, splitter="random"

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| BLACK VOICES | 0.25 | 0.24 | 0.25 | 392 |
| BUSINESS | 0.20 | 0.20 | 0.20 | 380 |
| COLLEGE | 0.43 | 0.38 | 0.40 | 212 |
| COMEDY | 0.36 | 0.34 | 0.35 | 399 |
| CRIME | 0.34 | 0.40 | 0.36 | 409 |
| CULTURE & ARTS | 0.24 | 0.41 | 0.30 | 673 |
| DIVORCE | 0.60 | 0.53 | 0.56 | 419 |
| EDUCATION | 0.36 | 0.32 | 0.34 | 198 |
| ENTERTAINMENT | 0.19 | 0.22 | 0.20 | 387 |
| ENVIRONMENT | 0.24 | 0.34 | 0.28 | 355 |
| FIFTY | 0.11 | 0.15 | 0.13 | 273 |
| GOOD NEWS | 0.24 | 0.15 | 0.18 | 305 |
| HEALTHY LIVING | 0.14 | 0.18 | 0.16 | 386 |
| HOME & LIVING | 0.50 | 0.51 | 0.50 | 386 |
| IMPACT | 0.14 | 0.17 | 0.15 | 400 |
| MEDIA | 0.45 | 0.37 | 0.41 | 395 |
| MONEY | 0.29 | 0.29 | 0.29 | 324 |
| PARENTING | 0.32 | 0.30 | 0.31 | 391 |
| POLITICS | 0.32 | 0.34 | 0.33 | 420 |
| QUEER VOICES | 0.75 | 0.55 | 0.63 | 415 |
| RELIGION | 0.50 | 0.37 | 0.43 | 440 |
| SCIENCE | 0.43 | 0.37 | 0.40 | 414 |
| SPORTS | 0.42 | 0.37 | 0.40 | 410 |
| STYLE | 0.35 | 0.33 | 0.34 | 413 |
| STYLE & BEAUTY | 0.52 | 0.38 | 0.44 | 391 |
| TASTE | 0.44 | 0.39 | 0.41 | 397 |
| TECH | 0.52 | 0.44 | 0.48 | 416 |
| TRAVEL | 0.27 | 0.30 | 0.28 | 405 |
| WEDDINGS | 0.73 | 0.64 | 0.68 | 400 |
| WEIRD NEWS | 0.19 | 0.14 | 0.17 | 408 |
| WELLNESS | 0.17 | 0.21 | 0.19 | 407 |
| WOMEN | 0.31 | 0.27 | 0.29 | 400 |
| WORLD | 0.39 | 0.31 | 0.34 | 404 |
| accuracy | | | 0.34 | 12824 |
| macro avg | 0.35 | 0.33 | 0.34 | 12824 |
| weighted avg | 0.36 | 0.34 | 0.34 | 12824 |

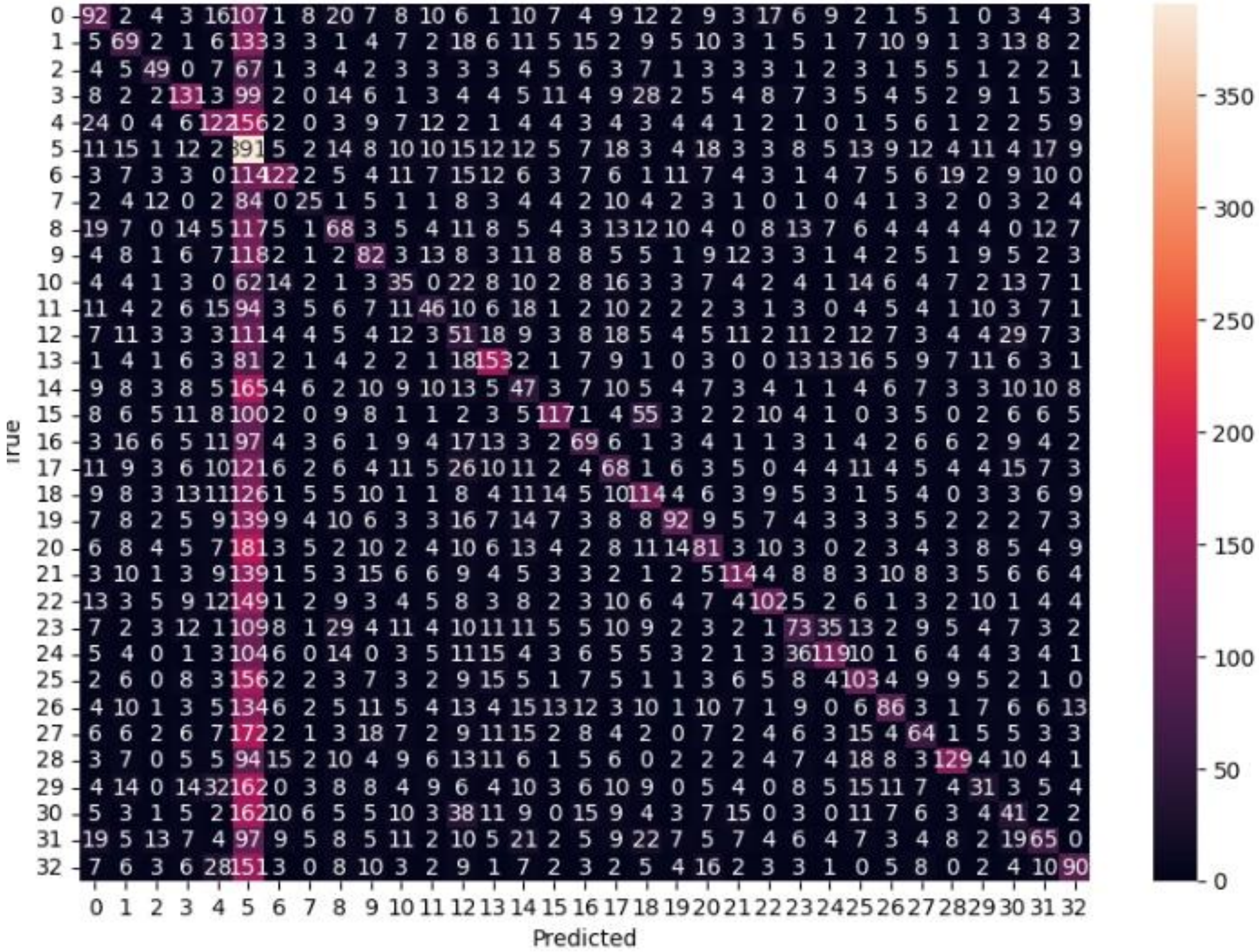
## tf-idf (unigram)



## min_samples_split=50, max_features=None, min_samples_leaf=5, splitter="random"

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| BLACK VOICES | 0.28 | 0.23 | 0.26 | 392 |
| BUSINESS | 0.25 | 0.18 | 0.21 | 380 |
| COLLEGE | 0.35 | 0.23 | 0.28 | 212 |
| COMEDY | 0.40 | 0.33 | 0.36 | 399 |
| CRIME | 0.34 | 0.30 | 0.32 | 409 |
| CULTURE & ARTS | 0.09 | 0.58 | 0.16 | 673 |
| DIVORCE | 0.47 | 0.29 | 0.36 | 419 |
| EDUCATION | 0.23 | 0.13 | 0.16 | 198 |
| ENTERTAINMENT | 0.23 | 0.18 | 0.20 | 387 |
| ENVIRONMENT | 0.29 | 0.23 | 0.26 | 355 |
| FIFTY | 0.15 | 0.13 | 0.14 | 273 |
| GOOD NEWS | 0.24 | 0.15 | 0.18 | 305 |
| HEALTHY LIVING | 0.12 | 0.13 | 0.13 | 386 |
| HOME & LIVING | 0.40 | 0.40 | 0.40 | 386 |
| IMPACT | 0.14 | 0.12 | 0.13 | 400 |
| MEDIA | 0.47 | 0.30 | 0.36 | 395 |
| MONEY | 0.27 | 0.21 | 0.24 | 324 |
| PARENTING | 0.21 | 0.17 | 0.19 | 391 |
| POLITICS | 0.31 | 0.27 | 0.29 | 420 |
| QUEER VOICES | 0.45 | 0.22 | 0.30 | 415 |
| RELIGION | 0.30 | 0.18 | 0.23 | 440 |
| SCIENCE | 0.47 | 0.28 | 0.35 | 414 |
| SPORTS | 0.45 | 0.25 | 0.32 | 410 |
| STYLE | 0.27 | 0.18 | 0.21 | 413 |
| STYLE & BEAUTY | 0.48 | 0.30 | 0.37 | 391 |
| TASTE | 0.31 | 0.26 | 0.28 | 397 |
| TECH | 0.37 | 0.21 | 0.27 | 416 |
| TRAVEL | 0.26 | 0.16 | 0.20 | 405 |
| WEDDINGS | 0.52 | 0.32 | 0.40 | 400 |
| WEIRD NEWS | 0.18 | 0.08 | 0.11 | 408 |
| WELLNESS | 0.16 | 0.10 | 0.12 | 407 |
| WOMEN | 0.26 | 0.16 | 0.20 | 400 |
| WORLD | 0.43 | 0.22 | 0.29 | 404 |
| accuracy | | | 0.24 | 12824 |
| macro avg | 0.31 | 0.23 | 0.25 | 12824 |
| weighted avg | 0.31 | 0.24 | 0.25 | 12824 |

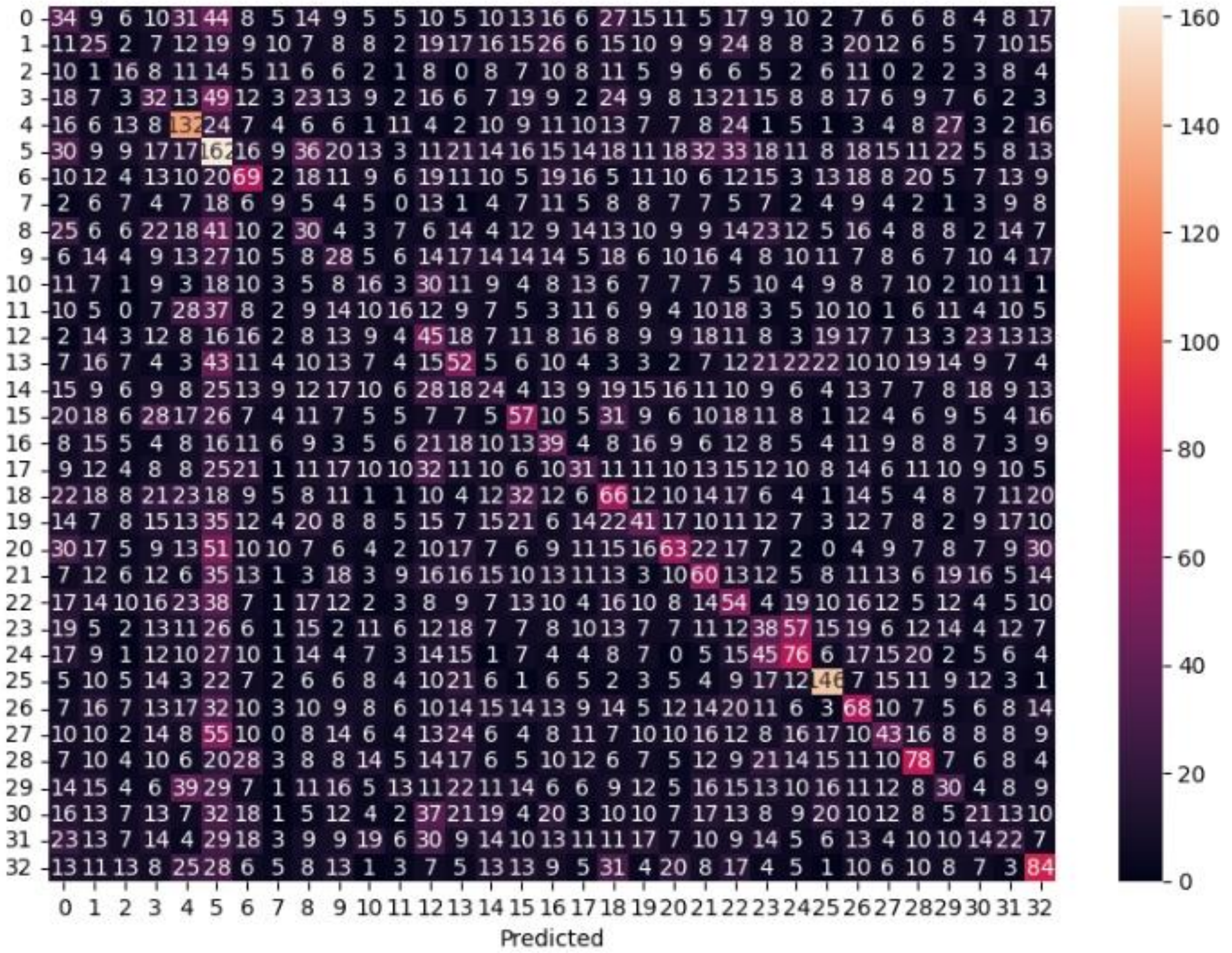
## tf-idf (bigram)



## min_samples_split=50, max_features=None, min_samples_leaf=5, splitter="random"

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| BLACK VOICES | 0.07 | 0.09 | 0.08 | 392 |
| BUSINESS | 0.07 | 0.07 | 0.07 | 380 |
| COLLEGE | 0.08 | 0.08 | 0.08 | 212 |
| COMEDY | 0.08 | 0.08 | 0.08 | 399 |
| CRIME | 0.24 | 0.32 | 0.27 | 409 |
| CULTURE & ARTS | 0.15 | 0.24 | 0.18 | 673 |
| DIVORCE | 0.16 | 0.16 | 0.16 | 419 |
| EDUCATION | 0.07 | 0.05 | 0.05 | 198 |
| ENTERTAINMENT | 0.08 | 0.08 | 0.08 | 387 |
| ENVIRONMENT | 0.08 | 0.08 | 0.08 | 355 |
| FIFTY | 0.07 | 0.06 | 0.06 | 273 |
| GOOD NEWS | 0.09 | 0.05 | 0.07 | 305 |
| HEALTHY LIVING | 0.09 | 0.12 | 0.10 | 386 |
| HOME & LIVING | 0.11 | 0.13 | 0.12 | 386 |
| IMPACT | 0.07 | 0.06 | 0.07 | 400 |
| MEDIA | 0.15 | 0.14 | 0.15 | 395 |
| MONEY | 0.10 | 0.12 | 0.11 | 324 |
| PARENTING | 0.10 | 0.08 | 0.09 | 391 |
| POLITICS | 0.14 | 0.16 | 0.15 | 420 |
| QUEER VOICES | 0.12 | 0.10 | 0.11 | 415 |
| RELIGION | 0.18 | 0.14 | 0.16 | 440 |
| SCIENCE | 0.14 | 0.14 | 0.14 | 414 |
| SPORTS | 0.11 | 0.13 | 0.12 | 410 |
| STYLE | 0.09 | 0.09 | 0.09 | 413 |
| STYLE & BEAUTY | 0.20 | 0.19 | 0.20 | 391 |
| TASTE | 0.36 | 0.37 | 0.36 | 397 |
| TECH | 0.15 | 0.16 | 0.16 | 416 |
| TRAVEL | 0.14 | 0.11 | 0.12 | 405 |
| WEDDINGS | 0.21 | 0.20 | 0.20 | 400 |
| WEIRD NEWS | 0.10 | 0.07 | 0.08 | 408 |
| WELLNESS | 0.08 | 0.05 | 0.06 | 407 |
| WOMEN | 0.08 | 0.06 | 0.06 | 400 |
| WORLD | 0.21 | 0.21 | 0.21 | 404 |
| accuracy | | | 0.13 | 12824 |
| macro avg | 0.13 | 0.13 | 0.13 | 12824 |
| weighted avg | 0.13 | 0.13 | 0.13 | 12824 |

## word2vec



# C o n c l u s i o n

1

TF-IDF unigrams generally have better results compared to TF-IDF bigrams and Word2Vec.

Bigrams focus on word pairs, which can miss out the importance of individual important words. And not all word pairs are meaningful and relevant, they can add noise.

TF-IDF highlights words that are unique to a document, helping to distinguish between different topics or categories, Word2Vec might miss the specific importance of words in  documents.

2

Poor performance in Multi-Class Classification

More classes increase the complexity of the decision boundaries the model has to learn, making it harder to distinguish between them.

# Literature

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.

Priyanka, & Kumar, D. (2020). Decision tree classifier: a detailed survey. International Journal of Information and Decision Sciences, 12(3), 246-269.

Quinlan, J. R. (1996). Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28(1), 71-72.

Rastogi, R., & Shim, K. (2000). PUBLIC: A decision tree classifier that integrates building and pruning. Data Mining and Knowledge Discovery, 4, 315-344.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28.

# THANK YOU