Decision Trees

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

CONTENTS

01 Introduction to Decision Trees **02** Concepts of Decision Trees **03** Advantages and disadvantages **04** Results



Introduction

01 02 03

ADD YOUR TITLE HERE

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.

Evolution and Adoption

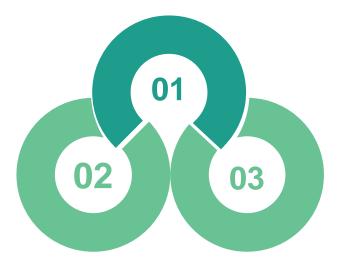
Over the years, decision trees have evolved greatly, with advancements in algorithms to handle overfitting, better ways to split nodes, and integration with other machine learning techniques like ensemble methods.

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

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.



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.

C4.5

M i l e s t o n e s

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.

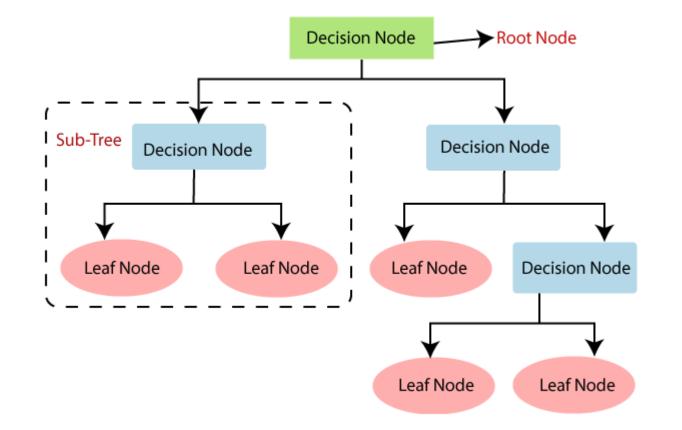
CART (Classification and Regression Trees)

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.



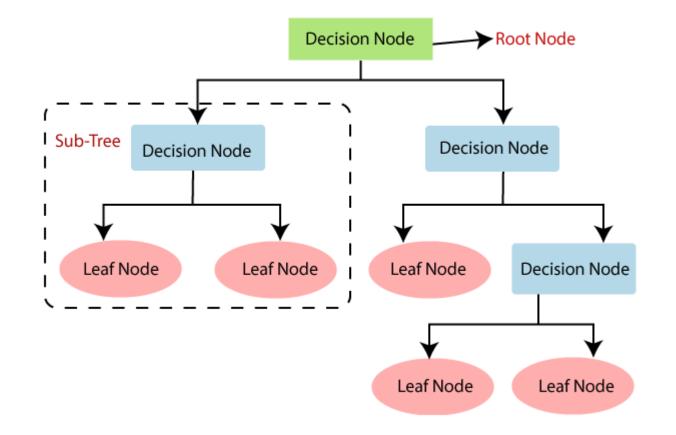
Concepts of Decision trees

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).

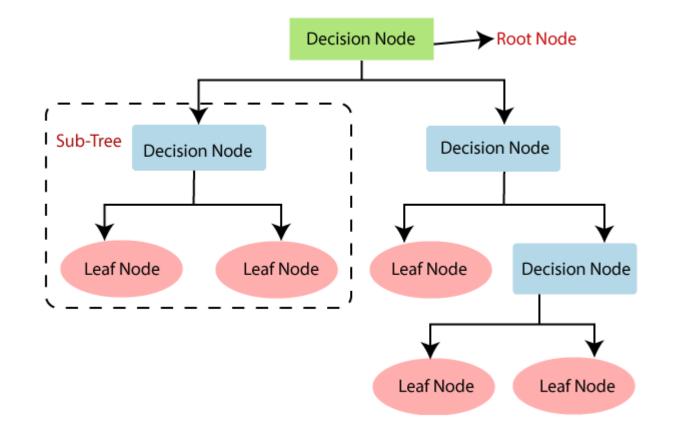
Step 3: Divide the root nide into subsets that contains possible values for the best attributes.

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.

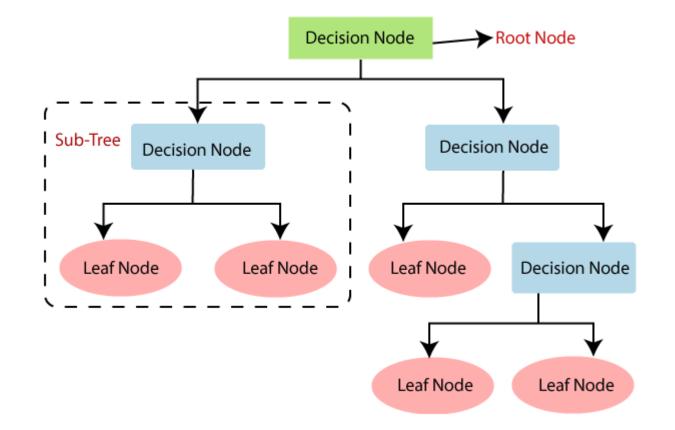
Information Gain is the reduction in entropy obtained by dividing the dataset according to a particular feature.

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.

Pruning



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.



Advantages and Disadvantages

A d v a n t a g e s

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.



Flexibility:

Decision trees can be used for both classification and regression tasks. Their performance can be enhanced through ensemble methods like Random Forests.



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.



Biased Trees with Imbalanced Datasets:

Decision trees can be biased toward the dominant class.

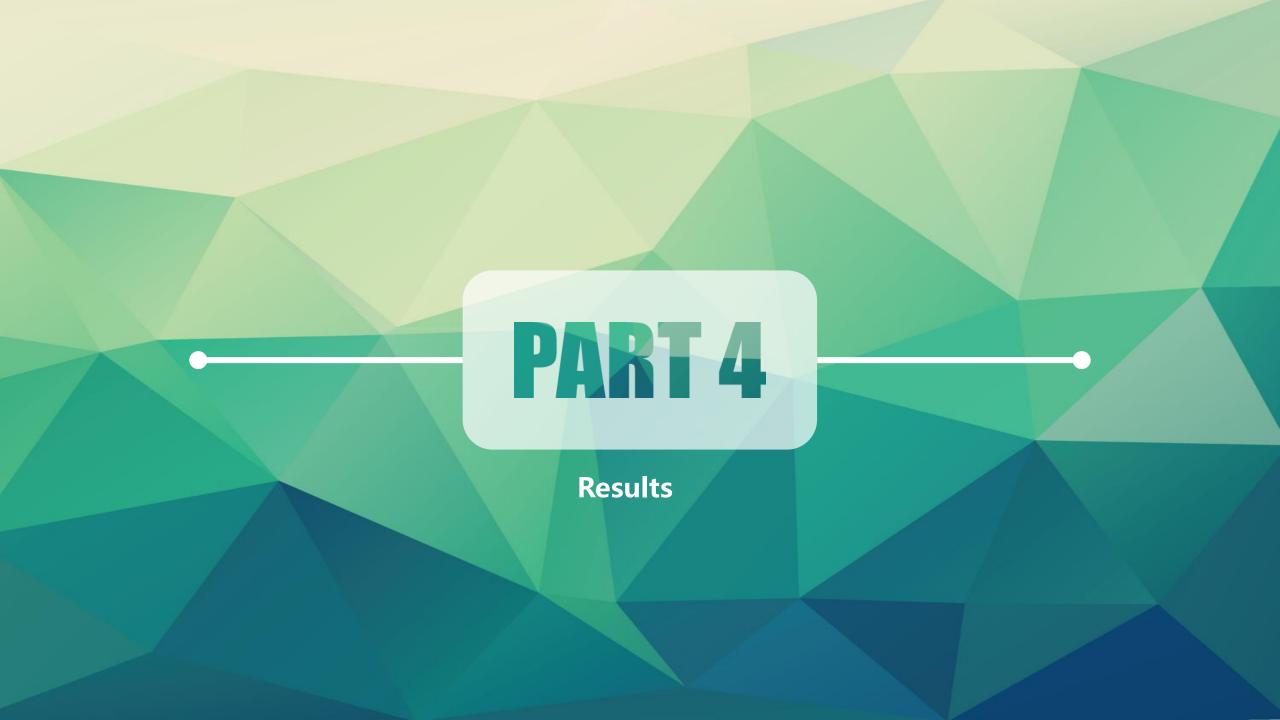
4

Difficulty in Capturing Complex Relationships:

While good for simple tasks, decision trees can struggle with tasks requiring the modeling of more complex relationships.

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.



parameters

max_depth:

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

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.

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.



max_features:

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

criterion:

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

splitter:

The strategy for choosing the split at each node.

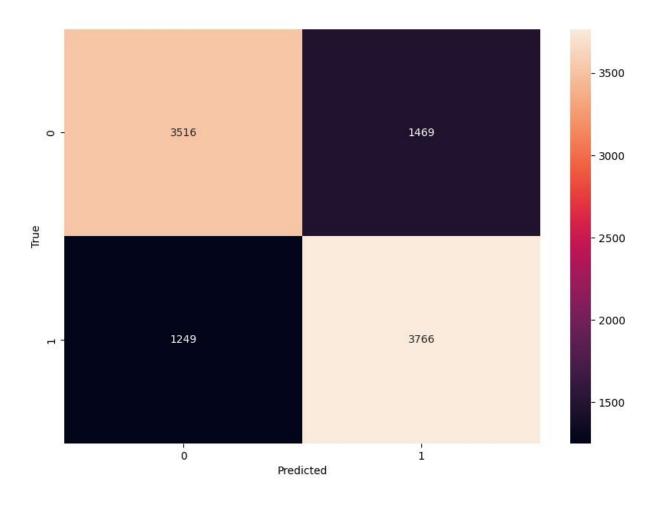
m H n

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



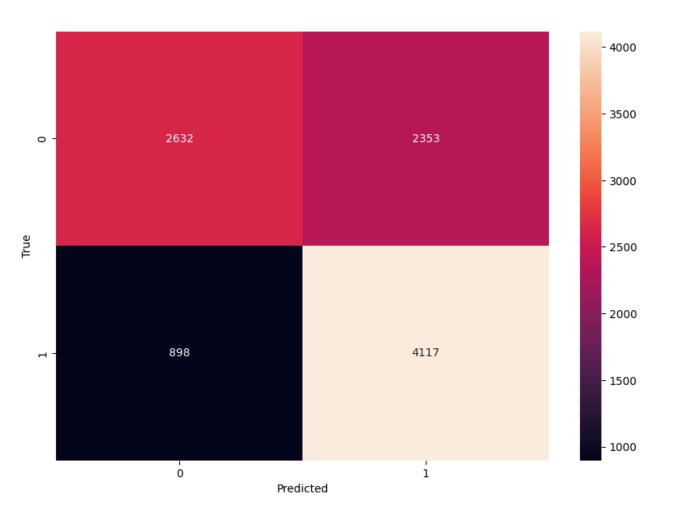
m H n

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

tf- idf(bigram)

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

Accuracy: 0.67 Macro Average:0.67 Weighted Average:0.67



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

m

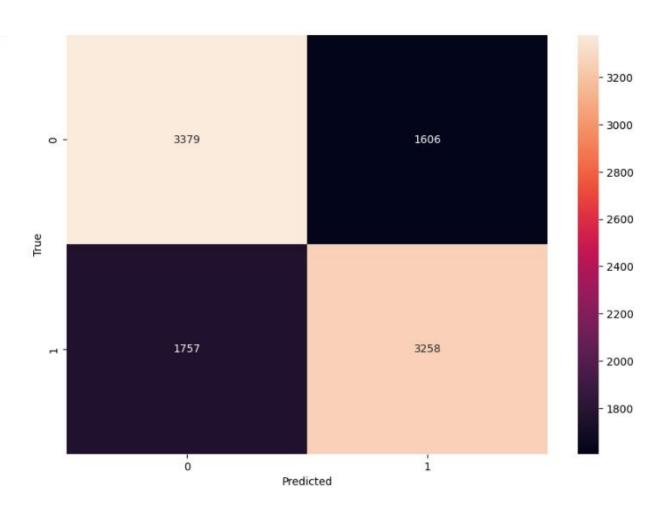
H

n

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 [a u t

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

Macro Average:0.67 Weighted Average:0.80 Accuracy: 0.80 tf-idf (unigram) 115 31 37 recall f1-score 0 0 2 precision support 0 -95 - 1600 Franz Kafka 0.60 0.41 0.49 280 Friedrich Schiller 0.54 0.38 0.45 266 - 1400 102 104 - -15 0 0 1 44 Henrik Ibsen 1.00 0.97 0.98 897 James Joyce 0.73 0.64 0.68 682 - 1200 0.39 Johann Wolfgang von Goethe 0.57 0.47 228 872 2 11 5 0 1 6 ~ -Virginia Woolf 0.87 0.92 0.89 1901 Wilhelm Busch 0.72 0.80 0.76 627 - 1000 True 3 438 0 0 0 0 244 0 4881 0.80 accuracy - 800 0.67 4881 0.69 0.67 macro avg weighted avg 0.80 0.80 0.80 4881 35 0 0 130 0 14 49 4 - 600 - 400 155 0 0 1 0 1745 0 <u>ہ</u> - 200 22 10 43 49 500 9 -0 i 2 0 3 4 5 6 Predicted

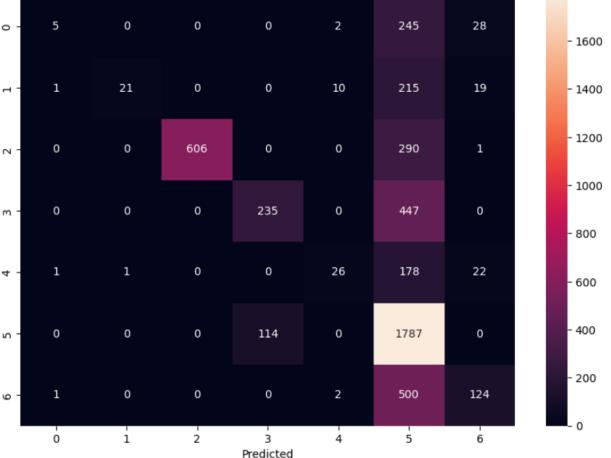
min_samples_split=20, max_features=None, min_samples_leaf=5, max_depth=30

A

Accuracy: 0.80

tf-idf (bigram)

recall f1-score precision support 0 0 5 0 -Franz Kafka 0.62 0.02 0.03 280 Friedrich Schiller 0.95 0.08 0.15 266 21 0 1 897 - -Henrik Ibsen 1.00 0.68 0.81 James Joyce 0.67 0.34 0.46 682 228 ohann Wolfgang von Goethe 0.65 0.11 0.19 606 0 0 2 1901 Virginia Woolf 0.49 0.94 0.64 627 Wilhelm Busch 0.64 0.20 0.30 True 3 0 0 0 4881 0.57 accuracy 0.37 4881 macro avg 0.72 0.34 weighted avg 0.67 0.57 0.52 4881 1 0 1 4 -



Macro Average:0.67 Weighted Average:0.80

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

recall f1-score precision support 0. Franz Kafka 0.32 0.36 0.34 280 Friedrich Schiller 0.35 0.35 0.36 266 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 N. Wilhelm Busch 0.61 0.60 0.61 627 True 3 0.70 accuracy 4881 0.55 macro avg 0.55 0.55 4881 weighted avg 0.71 0.70 0.70 4881

Word2Vec

101 43 3 0 35 0 98 - 1400 95 1 62 62 45 1 0 - 1200 848 8 2 21 10 1 - 1000 0 5 366 311 0 0 0 - 800 600 65 52 60 49 0 0 4 400 0 1 301 1599 0 0 5 0 - 200 62 24 108 52 0 378 9 0 1 2 4 5 6 3

Predicted

Accuracy: 0.70

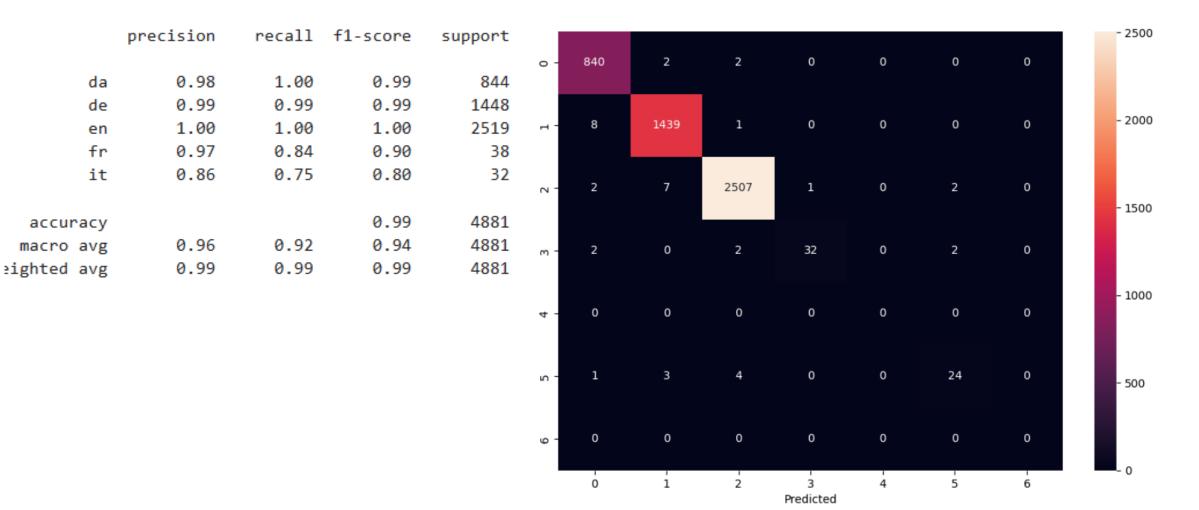
A

Macro Average: 0.55 Weighted Average: 0.70

Letter (language)

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

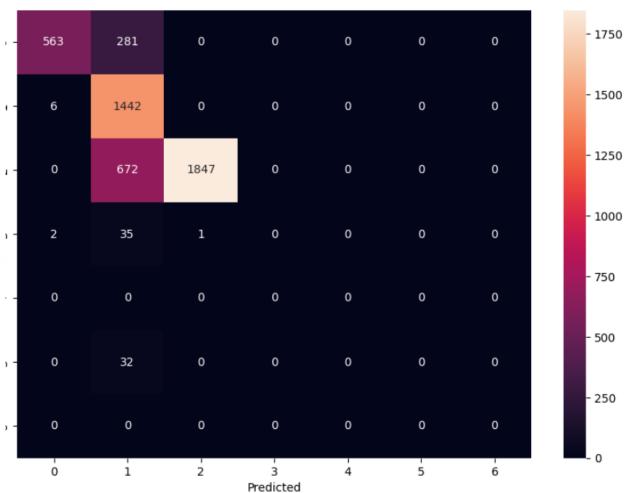
min_samples_split=10, max_features=None, min_samples_leaf=5, max_depth=30

precision recall f1-score support 563 281 1 -0.99 0.67 0.80 844 0.59 1.00 0.74 1448 1442 6 1 -1.00 0.73 0.85 2519 0.00 0.00 0.00 38 32 0.00 0.00 0.00 672 0 1 -0.79 4881 ۰-۱ 0.51 0.48 4881 35 0.48 2 0.86 0.79 0.79 4881 0 0

Accuracy: 0.79

A

Macro Average: 0.48 Weighted Average: 0.79



tf-idf (bigram)

accuracy macro avg

eighted avg

da de

en fr

it

etter

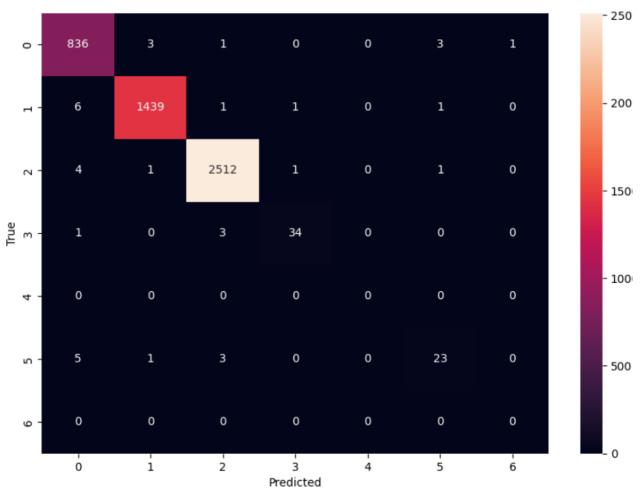
min_samples_split=10, max_features="sqrt"

Word2Vec

	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

Accuracy: 0.99

Macro Average: 0.78 Weighted Average: 0.99



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

e w

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)

S

	0	96	Л	5	6	20	26	2	10	25	10	5	2	10	1	٩	12	2	6	15	Л	10	1	16	7	6	1	Q	6	5	6	Q	12	2	
																																	14		
																													11				10		
																													10		8	7	7	2	
							219																								-	2	4 1	2	
							278																					-		_	_	_	19		
		6					15																							_			10	· · ·	
		3					15																					3		3		9	3	3	
	-						45																				_	3					101	10	
							27																										1 1		
	10 -																													_			14		
	11 -																													_	_	_	_	-	
	12 -																																		
	13 -																																		
	14 -																																		
	15 -	15	5	4	19	8	24	1	2	12	13	2	2	3	2	6	145	54	6	44	4	5	3	9	7	1	2	10	9	0	11	3	2 1	12	
True	16 -																												12	4	8	15	1	6	
F	17 -	8	8	4	7	8	24	9	8	13	7	12	12	15	5	19	1	6	119	96	2	4	3	3	12	8	6	7	8	1	2	29	13	2	
	18 -	6	19	1	12	16	32	3	6	12	20	8	2	9	1	9	22	12	6	141	.6	11	0	4	6	2	4	6	4	1	3	6	121	18	
	19 -	10	3	3	7	5	26	4	5	12	5	5	3	8	3	7	5	6	7	11	227	7	3	6	4	1	3	4	3	5	2	5	7	3	
	20 -	21	11	4	8	5	36	1	3	10	12	8	1	19	4	17	6	3	11	15	11	163	6	5	3	1	5	2	9	2	7	12	10	9	
	21 -	7	10	1	9	8	34	3	2	13	29	12	6	21	9	7	2	6	3	6	3	11	153	37	6	5	6	2	10	1	14	11	2	5	
	22 -	15	9	6	6	15	33	3	5	16	9	4	6	4	5	14	8	5	5	7	7	2	6	152	26	5	4	4	8	3	13	6	11	8	
	23 -																					2	3	4:	135	16	10	4	8	0	10	16	7	7	
	24 -																					2							26		3	6	5	1	
	25 -																																	2	
	26 -																																	7	
	27 -																											6	121	. 3	7	11	4	6	
	28 -																												11			8	5	1	
	29 -																																71		
	30 -																													1			6		
	31 -																									-	4	5	5	5			108		
	32 -	15	11	1	3	22	32	3	3	14	20	4	9	11						22			9	7	6	2	3	7	8	2	10	10	41	25	
		ò	i	ż	з	4	5	6	7	8	9	10	11	12									21	22	23	24	25	26	27	28	29	30	313	32	

- 250

- 200

- 150

- 100

- 50

- 0

Predicted

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

e w

	17-19-19-19-19-19-19-19-19-19-19-19-19-19-			000000000000000000000000000000000000000	
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	0
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	4
EDUCATION	0.23	0.13	0.16	198	1
ENTERTAINMENT	0.23	0.18	0.20	387	6
ENVIRONMENT	0.29	0.23	0.26	355	-
FIFTY	0.15	0.13	0.14	273	8
GOOD NEWS	0.24	0.15	0.18	305	5
HEALTHY LIVING	0.12	0.13	0.13	386	10
HOME & LIVING	0.40	0.40	0.40	386	11
IMPACT	0.14	0.12	0.13	400	12
MEDIA	0.47	0.30	0.36	395	13
MONEY	0.27	0.21	0.24	324	14
PARENTING	0.21	0.17	0.19	391	. 15
POLITICS	0.31	0.27	0.29	420	2 16
QUEER VOICES	0.45	0.22	0.30	415	= 17
RELIGION	0.30	0.18	0.23	440	18
SCIENCE	0.47	0.28	0.35	414	19
SPORTS	0.45	0.25	0.32	410	20
STYLE	0.27	0.18	0.21	413	21
STYLE & BEAUTY	0.48	0.30	0.37	391	22
TASTE	0.31	0.26	0.28	397	23
TECH	0.37	0.21	0.27	416	24
TRAVEL	0.26	0.16	0.20	405	25
WEDDINGS	0.52	0.32	0.40	400	
WEIRD NEWS	0.18	0.08	0.11	408	20
WELLNESS	0.16	0.10	0.12	407	
WOMEN	0.26	0.16	0.20	400	28
WORLD	0.43	0.22	0.29	404	29
accuracy			0.24	12824	31
macro avg	0.31	0.23	0.25	12824	32
weighted avg	0.31	0.24	0.25	12824	

recall f1-score support

precision

tf-idf (bigram)

S

1 2 3 4 5 6 7 8 9 10 11	- 8 2 -24 0 -11 1 - 3 7 - 2 4 -19 7 - 4 8 - 4 8 - 4 4 -11 4	9 2 49 2 4 5 1 3 12 0 1 1	1 0 131 61 12 3	61 7 2 2 01 2 51 71 0	071 333 671 992 915 142 840 175 182 521 943	3 0 2 22 25 1 4 2	20 1 4 14 5 1 68 2 1 6	4 2 6 9 8 4 5	7 3 1 7 10 11 1 5 3 35	10 2 3 12 10 7 1 4 13 0 46		6 3 4 12 12 3 8 3 8 3 8	10 11 4 5 4 12 6 4 5 11 10 18		15 6 4 3 7 7 2 3 8 8 2	9 2 3 9 4 18 6 10 13 5 16	12 9 7 28 3 1 4 12 5 3 2	2 5 1 2 4 4 11 2 10 1 3 2	9 3 5 4 18 7 3 4 9 7 2	3 3 4 1 3 4 1 0 12 4 3	17 1 3 8 2 3 3 0 8 3 2 1	6 5 1 7 1 8 1 1 3 4 3	9 2 3 0 5 4 0 7 1 1 0	2 7 3 5 1 3 7 4 6 4 14 4	1 10 1 4 5 1 4 2 6 5	5 9 5 6 12 6 3 4 5 4 4	1 5 2 1 19 2 4 1 7 1	0 3 1 9 2 1 2 0 4 9 2 1	3 13 2 1 2 4 9 5 13 3	8 2 5 5 17	3 2 1 3 9 9 0 4 7 3 1		350
23 24 25 26 27	- 1 4 - 9 8 - 3 10 - 11 9 - 9 8 - 7 7 7 7 - 7 7 8 - 7 7 7 7 - 7 7 7	3 5 6 3 2 4 0 1 5 3 0 1 5 0 0 1 2 0 0 1 2 0 0 1 2	3681156135539121836	51 81 101 111 91 71 91 121 31 31 51 71	1114 31 2 654 002 97 4 216 261 399 813 391 491 098 046 562 346 722	5452	5 4 2 9 6 6 5 10 2 3 9 29 4 3 5 3 1	4 2 10 8 1 4 10 6 10 15 3 4 0 7 11 18	12 9 1 9 1 1 3 2 6 4 11 3 3 5 7	3 10 1 4 5 1 3 4 6 5 4 5 2 4 2	181 13 2 17 26 8 16 9 8 10 9 13 9 13 9	18 5 3 13 10 4 7 6 4 3 11 15 4 11 15 4 11	47 51 3 11 14 13 5 8 11 4 5 15 15	3 1 3 1 2 2 1 4 7 4 3 2 5 3 1 1 3 2 2 1 4 7 4 3 2 5 3 1 1 3 2 5 3 1 1 3 2 5 3 1 1 3 2 5 3 1 3 1 3 2 5 3 1 3 1 3 2 5 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3	8 7 7 1 69 4 5 3 2 3 3 5 6 7 12 8	189104681018821015534	5 1 5 5 5 1 1 4 8 11 1 6 9 5 1 0 2	4 0 4 3 3 6 4 92 14 2 4 2 3 1 1 0	5 3 7 2 4 3 6 9 8 1 5 7 3 2 3 10 7	10 3 2 1 5 3 5 3 4 2 1 6 7 2 0	2 0 4 10 9 7 10 4 02 1 3 5 1 4	11 13 1 4 3 4 5 4 3 8 5 73 361 8 9 6	2 13 1 1 1 4 3 0 8 2 35 19 41 0 3	12 16 4 0 4 11 1 3 2 3 6 13 10 03 6 15	7 5 6 3 2 4 5 3 3 10 1 2 1 4 6 4 6 4	3 9 7 5 6 5 4 5 4 8 3 9 6 9 3 4	471 3064023321 491	411322432851044575	6 9 15 3 2 5 6 1 7 3 2 6 5	3 6 4 7 6 7 4 6 4 3 4 6 1 3	3185239394421033		200 150
28 29 30 31 32	-53	1 13		321	94 1: 62 0 6210 97 9 51 3 5 6	0	10 8 5 8 8 8	4 5 5 10 9	9 4 10 11 3 10	6 9 3 2 2	13 6 38 10 9	1	6 10 9 21 7 14	1 3 0 2 2 15	5 6 15 5 3 16	6 10 9 2 17	0 9 4 22 5	2 3 7 4	2 5 7 16 20	2 4 15 7 2	4 0 4 3	7 8 3 6 3 	4 5 4 1 24	18 15 11 7 0	1	31 7 6 4 8	4 3 8 0	2	3 41 19(4	4 5 2 55 10 9 31 3	0		50 0

Predicted

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

170 N	precision	recall	f1-score	supp <mark>o</mark> rt
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
ULTURE & 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

\$

	0	34 9	6	10	31	44	8	5	14	9	5	5	10	5	10	13	16	6	27	15	11	5	17	9	10	2	7	б	6	8	4	8 17		- 160
	1	1125	2	7	12	19	9	10	7	8	8	2	19	17	16	15	26	6	15	10	9	9	24	8	8	3	201	12	6	5	7	1015		
	100	10 1				_						-						8	_		9		1.		-	6	11	0	2	2	3	8 4		
		-18 7		-	_	Statement of the																	21			8		6	9	7	6	23		
		16 6									1							10					24				3	4	8 2	27	3	2 16		- 140
		-30 9		17		1000								_				14		_			33		_	- C	18		112	22	5	8 13		Sho Sha Sha
		10 12					COLUMN 2														10	6	-			13	18	8 2	20	5	201	13 9		
		2 6				18			and the second	4			13			7			8		7	7	5	7	2	4	9	4	2	1	3	98		
		25 6		_	_			-	-				6				100	-			. 7		14		_	5			2	8		14 7		- 120
	- T	6 14											14							6	10	16		1	10	11			6	1.0		4 17		
	_	11 7			3			-	-			-	30	_			-	13		7	7	7	-		4		8	5		2 1		11 1		
		10 5		7		-					_		12					_	6	-		10		-	2.1	10	10	1	6 1	1		10 5		
		2 14																			9	1000		~	-	19	17	7	13	3 2	3	1313		- 100
		7 16																										10 1	191	.4	9	14		1 annes
		15 9																							-		13			8 1	8	9 13		
e		2018																								1	12		6	9	5	4 16		L.
True		8 15				-							21	-			-						_			4		9.	8	8	/	3 9		- 80
		9 12																										-				10 5		
		2218													100						100							2		8	1	1120		
		-14 7																												2 8		1710		
	-	30 17																				COLUMN STATE						9		-		9 30 5 14		- 60
	_	-17 14																	100			1000			100		161			191		5 10		
		19 5			_	-			-										-			- 11	1000	Contract of the	_		-		100	2		12 7		
		17 9																					_								5	100		
		5 10																							and the second second	and and an	7				2	3 1		- 40
	26	and the second se																								2.00.2	Sector 1		7	5 1	6	9 14		
	_	10 10																											16	8	e e	8 9		
		7 10																											and a local division of the		6 6	8 4		
		14 15			-					16			11										15					2		30		8 9		- 20
	-	16 13														-	-		1.2								101		8	5 2	1	1310		
		2313		1000		and so its and	Contraction in which the		100	-			30							-		1000		100	-		13		-	01	4	22 7		
		13 11		-		-	-											_						-								3 84		
	32	11	-	1	-	-	1	Ĩ	1	1	1	1				1	1		1	-	1		1	1	1		1	1	1	1	1	1 1		-0
		0 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	223	23	242	25	262	272	282	93	0	31 32		
															F	Ter	dic	ter	1															

Predicted

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