Natural Language Processing

Project NLQ-Class

Elaborated by:

Ala Eddine Naily
Houssem Mallouli
Mohanned Annéne Barkallah
3ATEL DASEC

Supervised by :

Mr Mustapha BENHAJ MINIAOUI
Mme Linda MARRAKCHI

Plan

- Introduction
- DataSets
- Implementation
- Video Demonstarion
- Conclusion & Perspectives

Introduction

Problematic : We want to train a model to predict the category of any given question.

Proposed Solution

Use a dataset of questions and their corresponding categories and subcategories to train classification models to predict the categories and their labels.

DataSet

Provided Datasets : Training set

Training and Test sets

- Training set 1(1000 labeled questions)
- Training set 2(2000 labeled questions)
- Training set 3(3000 labeled questions)
- Training set 4(4000 labeled questions)
- Training set 5(5500 labeled questions)
- Test set: TREC 10 questions

Training Set Selection

- Training set 5(5500 labeled questions)

Raw Data

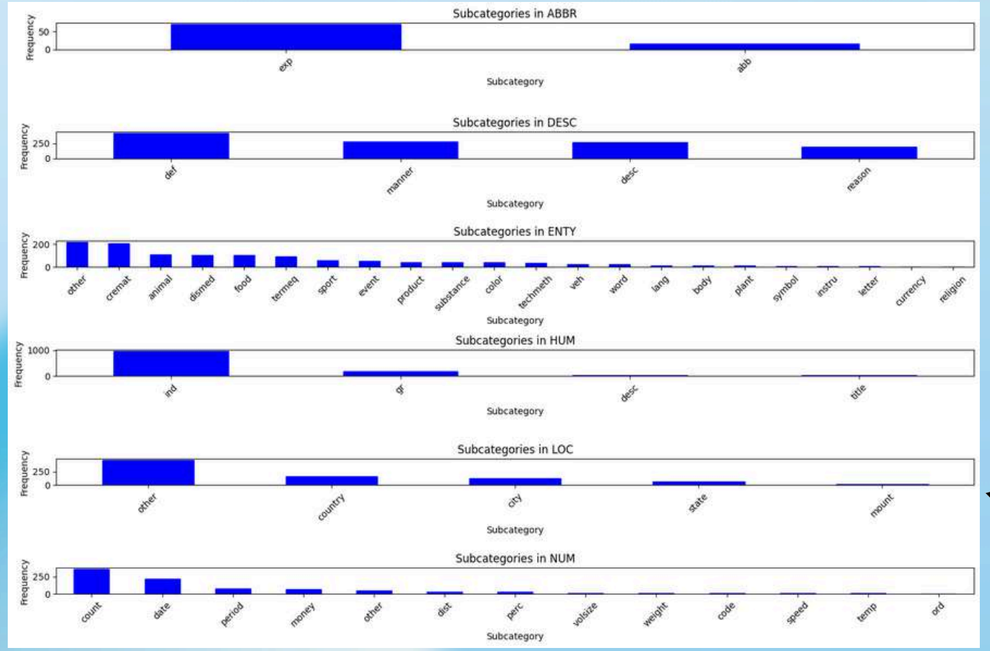
159 ENTY:cremat What poem contains the line , `` grow old with me the best is yet to be '' .

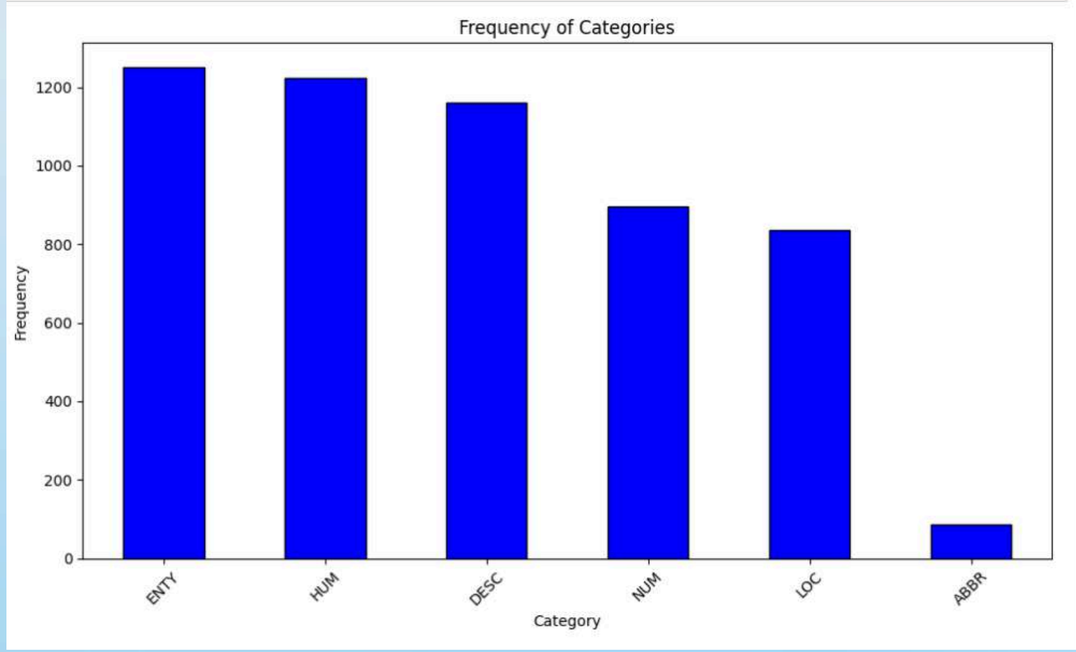
Dataset Reconstruction

1	Question	Category	Subcategory
2	How did serfdom develop in and then leave Russia ?	DESC	manner
3	What films featured the character Popeye Doyle ?	ENTY	cremat
4	How can I find a list of celebrities ' real names ?	DESC	manner
5	What fowl grabs the spotlight after the Chinese Year of the Monkey ?	ENTY	animal
6	What is the full form of .com ?	ABBR	exp
7	What contemptible scoundrel stole the cork from my lunch ?	HUM	ind
8	What team did baseball 's St. Louis Browns become ?	HUM	gr
9	What is the oldest profession ?	HUM	title

DataSet

Data Exploration : training set





5452 rows ×3 columns

DataSet

Provided Datasets : Testing set

Training and Test sets

- Training set 1(1000 labeled questions)
- Training set 2(2000 labeled questions)
- Training set 3(3000 labeled questions)
- Training set 4(4000 labeled questions)
- Training set 5(5500 labeled questions)
- Test set: TREC 10 questions

Test Set Selection

- Test set: TREC 10 questions

Raw Data

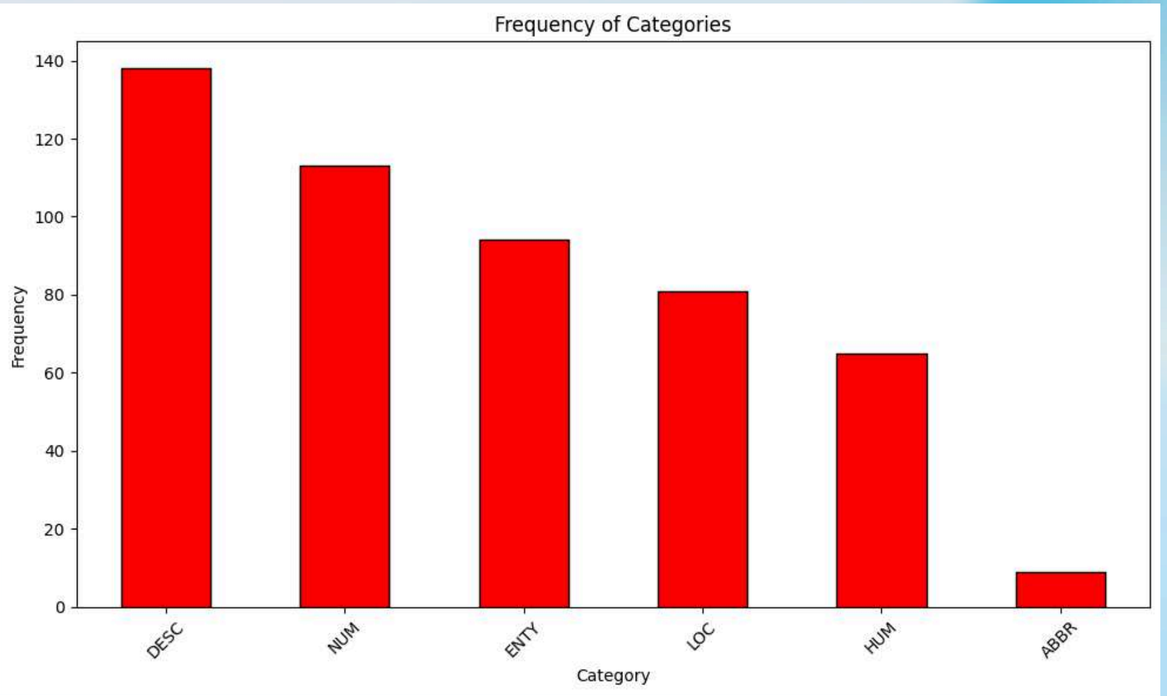
```
NUM:dist How far is it from
LOC:city What county is Mode
HUM:desc Who was Galileo ?
DESC:def What is an atom ?
```

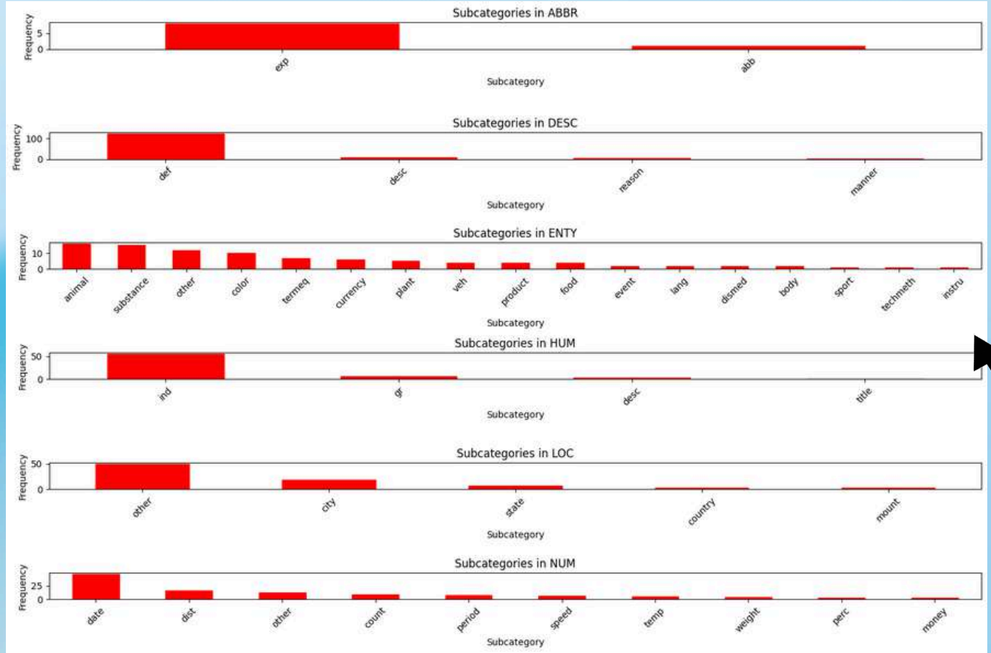
Dataset Reconstruction

1	Question,Category,Subcategory
2	How far is it from Denver to Aspen ?,NUM,dist
3	What county is Modesto , California in ?,LOC,city
4	Who was Galileo ?,HUM,desc
5	What is an atom ?,DESC,def
6	When did Hawaii become a state ?,NUM,date
7	How tall is the Sears Building ?,NUM,dist
8	George Bush purchased a small interest in which baseball team ?,HUM,gr
9	What is Australia 's national flower ?,ENTY,plant
10	Why does the moon turn orange ?,DESC,reason
11	What is autism ?,DESC,def
12	What city had a world fair in 1900 ?,LOC,city
13	What person 's head is on a dime ?,HUM,ind
14	What is the average weight of a Yellow Labrador ?,NUM,weight
15	Who was the first man to fly across the Pacific Ocean ?,HUM,ind
16	When did Idaho become a state ?,NUM,date
17	What is the life expectancy for crickets ?,NUM,other
18	What metal has the highest melting point ?,ENTY,substance
19	Who developed the vaccination against polio ?,HUM,ind
20	What is epilepsy ?,DESC,def
21	What year did the Titanic sink ?,NUM,date
22	Who was the first American to walk in space ?,HUM,ind
23	What is a biosphere ?,DESC,def
24	What river in the US is known as the Big Muddy ?,LOC,other
25	What is bipolar disorder ?,DESC,def
26	What is cholesterol ?,DESC,def
27	Who developed the Macintosh computer ?,HUM,ind

DataSet

Data Exploration : Testing set





500 rows ×3 columns

Implementation

- Lowercasing and punctuation Removal
- Stop-words removal and lemmatization with spaCy
- Text Vectorization: TF-IDF Feature Extraction
- First approach : combining the category and the sub category into one class
- Second approach : Predicting the subcategory based on the predicted category

Implementation

1.Lowercasing and punctuation Removal

1 What films featured the character Popeye Doyle ?

- str.lower() : Convert the text to lowercase
- str.replace(): Remove punctuation

1 what films featured the character popeye doyle

Implementation

2.Stop-words removal and lemmatization with spaCy

```
"what sprawling u s state boasts the most airports"
```

Stop-words removal and lemmatization with spaCy

```
what sprawl u s state boast airport
```

Natural Language Processing

Implementation

2.Stop-words removal and lemmatization with spaCy

Tokenization and lemmatization of the input text using spaCy

Removing predefined stop words from the text, except for WH-questions and the ones that are relevant to the question category.

Returning a cleaned-up string with significant words (lemmatized) only, which can be used for further processing

Implementation

3.Text Vectorization: TF-IDF Feature Extraction

Parameter	Explanation
ngram_range = (1, 2)	Generates both unigrams (1-word) and bigrams (2-words) from text.
max_features = 1000	Limits the model to the 1000 most frequent terms based on TF-IDF.
min_df = 5	Keeps terms that appear in at least 5 documents. Filters out rare terms.
max_df = 0.9	Ignores terms that appear in more than 90% of documents (common terms).
smooth_idf = True	Adds smoothing to avoid zero values in IDF.
sublinear_tf = True	Applies sublinear scaling to term frequency, reducing impact of frequent terms.

Implementation

4.First approach : combining the category and the sub category into one class

Category
DESC
ENTY
DESC
ENTY
ABBR

+

Subcategory
manner
cremat
manner
animal
exp

Combined_Label
DESC_manner
ENTY_cremat
DESC_manner
ENTY_animal
ABBR_exp

Natural Language Processing

Implementation

4.First approach : combining the category and the sub category into one class

SVM Model : (Linear Kernel)

Combined Label Classification Report:

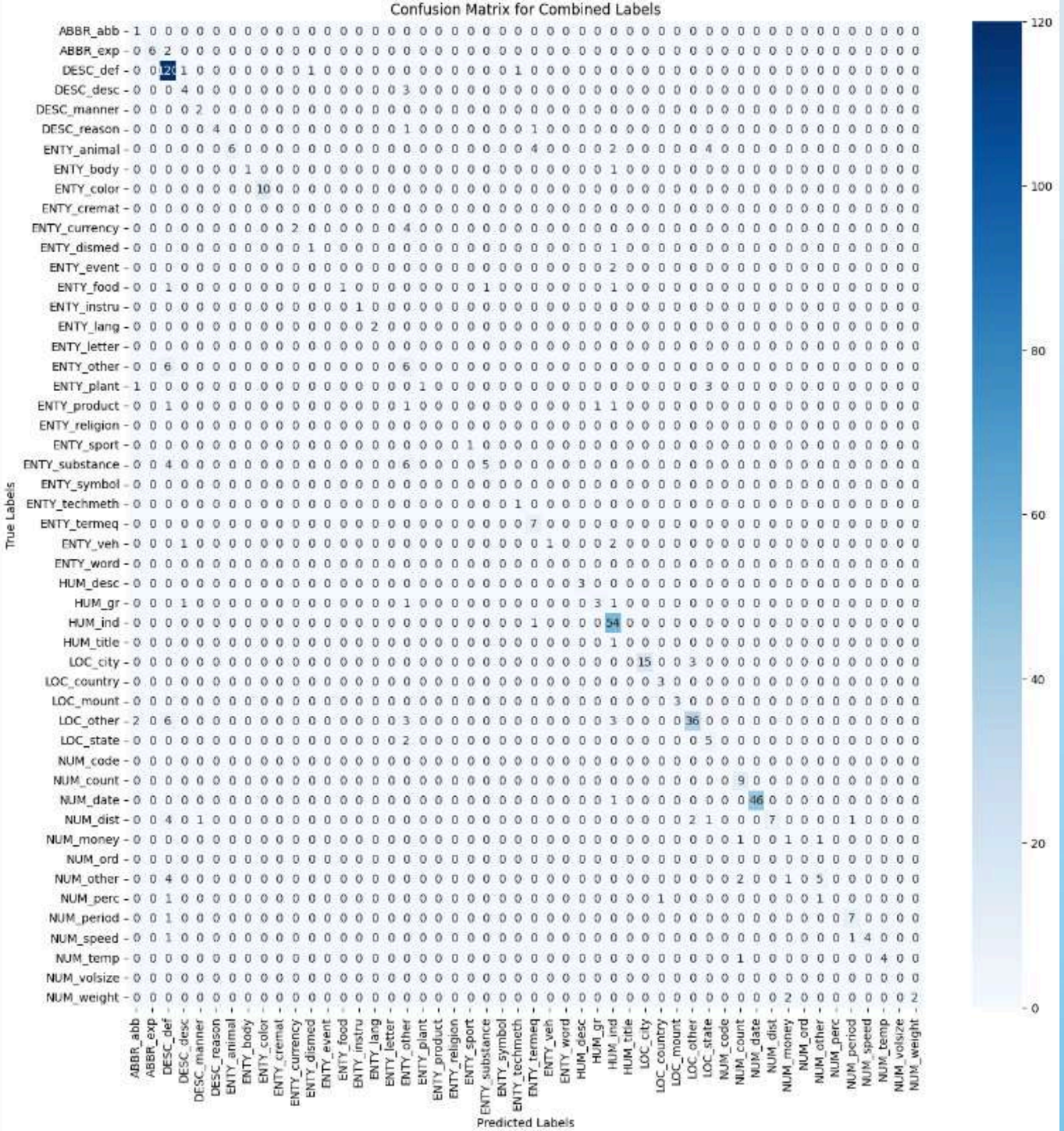
	precision	recall	f1-score	support
ABBR_abb	0.25	1.00	0.40	1
ABBR_exp	1.00	0.75	0.86	8
DESC_def	0.81	0.98	0.89	123
DESC_desc	0.67	0.57	0.62	7
DESC_manner	0.67	1.00	0.80	2
DESC_reason	1.00	0.67	0.80	6
ENTY_animal	1.00	0.38	0.55	16
ENTY_body	1.00	0.50	0.67	2
ENTY_color	1.00	1.00	1.00	10
ENTY_currency	1.00	0.33	0.50	6
ENTY_dismed	0.50	0.50	0.50	2
ENTY_event	0.00	0.00	0.00	2
ENTY_food	1.00	0.25	0.40	4
ENTY_instru	1.00	1.00	1.00	1
ENTY_lang	1.00	1.00	1.00	2
ENTY_other	0.21	0.50	0.29	12
ENTY_plant	1.00	0.20	0.33	5
ENTY_product	0.00	0.00	0.00	4
ENTY_sport	1.00	1.00	1.00	1
ENTY_substance	0.83	0.33	0.48	15
ENTY_techmeth	0.50	1.00	0.67	1
...				
accuracy			0.78	500
macro avg	0.74	0.64	0.64	500
weighted avg	0.82	0.78	0.77	500

Accuracy : 78%

Macro average for recall : 64%

Macro average for F1-score : 64%

Macro average for precision : 74%


Confusion Matrix for Combined Labels
True Labels
Predicted Labels
ABBR_abb
ABBR_exp
DESC_def
DESC_desc
DESC_manner
DESC_reason
ENTY_animal
ENTY_body
ENTY_color
ENTY_cremat
ENTY_currency
ENTY_dismed
ENTY_event
ENTY_food
ENTY_instru
ENTY_lang
ENTY_letter
ENTY_other
ENTY_plant
ENTY_product
ENTY_religion
ENTY_sport
ENTY_substance
ENTY_symbol
ENTY_techmeth
ENTY_termeq
ENTY_veh
ENTY_word
HUM_desc
HUM_gr
HUM_ind
HUM_title
LOC_city
LOC_country
LOC_mount
LOC_other
LOC_state
NUM_code
NUM_count
NUM_date
NUM_dist
NUM_money
NUM_ord
NUM_other
NUM_perc
NUM_period
NUM_speed
NUM_temp
NUM_volsize
NUM_weight
120
100
80
60
40
20
0

Natural Language Processing

Implementation

4.First approach : combining the category and the sub category into one class

Random Forest Classifier :

Combined Label Classification Report:

	precision	recall	f1-score	support
ABBR_abb	1.00	1.00	1.00	1
ABBR_exp	1.00	0.75	0.86	8
DESC_def	0.77	0.98	0.86	123
DESC_desc	0.75	0.43	0.55	7
DESC_manner	0.67	1.00	0.80	2
DESC_reason	1.00	0.83	0.91	6
ENTY_animal	1.00	0.38	0.55	16
ENTY_body	0.00	0.00	0.00	2
ENTY_color	1.00	0.90	0.95	10
ENTY_currency	0.00	0.00	0.00	6
ENTY_dismed	0.50	0.50	0.50	2
ENTY_event	0.00	0.00	0.00	2
ENTY_food	1.00	0.25	0.40	4
ENTY_instru	1.00	1.00	1.00	1
ENTY_lang	1.00	1.00	1.00	2
ENTY_other	0.21	0.42	0.28	12
ENTY_plant	0.00	0.00	0.00	5
ENTY_product	0.00	0.00	0.00	4
ENTY_sport	1.00	1.00	1.00	1
ENTY_substance	0.83	0.33	0.48	15
ENTY_techmeth	0.33	1.00	0.50	1
ENTY_termeq	0.78	1.00	0.88	7
...				
accuracy			0.75	500
macro avg	0.63	0.53	0.53	500
weighted avg	0.75	0.75	0.72	500

Accuracy : 75%

Macro average for recall : 53%

Macro average for F1-score : 53%

Macro average for precision : 63%

Natural Language Processing

Implementation

4.First approach : combining the category and the sub category into one class

Gradient Boosting Classifier :

Combined Label Classification Report:

	precision	recall	f1-score	support
ABBR_abb	0.33	1.00	0.50	1
ABBR_exp	1.00	0.75	0.86	8
DESC_def	0.81	0.96	0.88	123
DESC_desc	0.33	0.43	0.38	7
DESC_manner	1.00	0.50	0.67	2
DESC_reason	1.00	0.83	0.91	6
ENTY_animal	0.67	0.50	0.57	16
ENTY_body	0.00	0.00	0.00	2
ENTY_color	1.00	1.00	1.00	10
ENTY_currency	0.00	0.00	0.00	6
ENTY_dismed	0.50	0.50	0.50	2
ENTY_event	0.33	0.50	0.40	2
ENTY_food	0.50	0.25	0.33	4
ENTY_instru	1.00	1.00	1.00	1
ENTY_lang	1.00	0.50	0.67	2
ENTY_other	0.43	0.25	0.32	12
ENTY_plant	1.00	0.20	0.33	5
ENTY_product	0.00	0.00	0.00	4
ENTY_sport	1.00	1.00	1.00	1
ENTY_substance	0.55	0.40	0.46	15
ENTY_techmeth	0.25	1.00	0.40	1
ENTY_termeq	0.64	1.00	0.78	7
...				
accuracy			0.75	500
macro avg	0.65	0.58	0.58	500
weighted avg	0.76	0.75	0.73	500

Accuracy : 75%

Macro average for recall : 58%

Macro average for F1-score : 58%

Macro average for precision : 65%

Implementation

4.First approach : combining the category and the sub category into one class

Combined Models :

- Create an Ensemble Model using VotingClassifier.
- Combine Multiple Models: SVC, Random Forest, Logistic Regression, Naive Bayes, K-Nearest Neighbors, and Gradient Boosting.
- Set Voting Strategy: Use 'hard' voting for majority class voting.

- Enhance Robustness by combining predictions to improve accuracy and reduce overfitting.
- Aggregate Predictions by using the majority decision from all models for final classification.

Natural Language Processing

Implementation

4.First approach : combining the category and the sub category into one class

Combined Models :

	precision	recall	f1-score	support
ABBR_abb	0.33	1.00	0.50	1
ABBR_exp	1.00	0.75	0.86	8
DESC_def	0.72	0.99	0.84	123
DESC_desc	0.86	0.86	0.86	7
DESC_manner	0.67	1.00	0.80	2
DESC_reason	1.00	0.83	0.91	6
ENTY_animal	1.00	0.38	0.55	16
ENTY_body	0.00	0.00	0.00	2
ENTY_color	1.00	1.00	1.00	10
ENTY_currency	1.00	0.33	0.50	6
ENTY_dismed	0.50	0.50	0.50	2
ENTY_event	0.00	0.00	0.00	2
ENTY_food	1.00	0.25	0.40	4
ENTY_instru	1.00	1.00	1.00	1
ENTY_lang	1.00	0.50	0.67	2
ENTY_other	0.33	0.42	0.37	12
ENTY_plant	1.00	0.20	0.33	5
ENTY_product	0.00	0.00	0.00	4
ENTY_sport	1.00	1.00	1.00	1
ENTY_substance	0.83	0.33	0.48	15
ENTY_techmeth	0.50	1.00	0.67	1
ENTY_termeq	0.64	1.00	0.78	7
ENTY_veh	1.00	0.25	0.40	4
...				
accuracy			0.77	500
macro avg	0.76	0.60	0.62	500
weighted avg	0.80	0.77	0.75	500

Accuracy : 77%

Macro average for recall : 60%

Macro average for F1-score : 62%

Macro average for precision : 76%

Natural Language Processing

Implementation

4.First approach : combining the category and the sub category into one class

ANN Model : (15 epochs)

- Input layers : 512 units , relu activation
- Dropout layer : 0,3
- Hidden layer : 256 units, relu activation
- Dropout layer : 0,3
- Output layers: *Num of sub-categories* classes, Softmax activation

```
86/86 ━━━━━━━━━━━━━━━━━━━━ 1s 7ms/step - accuracy: 0.9355 - loss: 0.2551 - val_accuracy: 0.7960 - val_loss: 1.0151
...
     accuracy                           0.81       500
    macro avg       0.71      0.67      0.66       500
 weighted avg       0.82      0.81      0.80       500
```

Accuracy : 81%

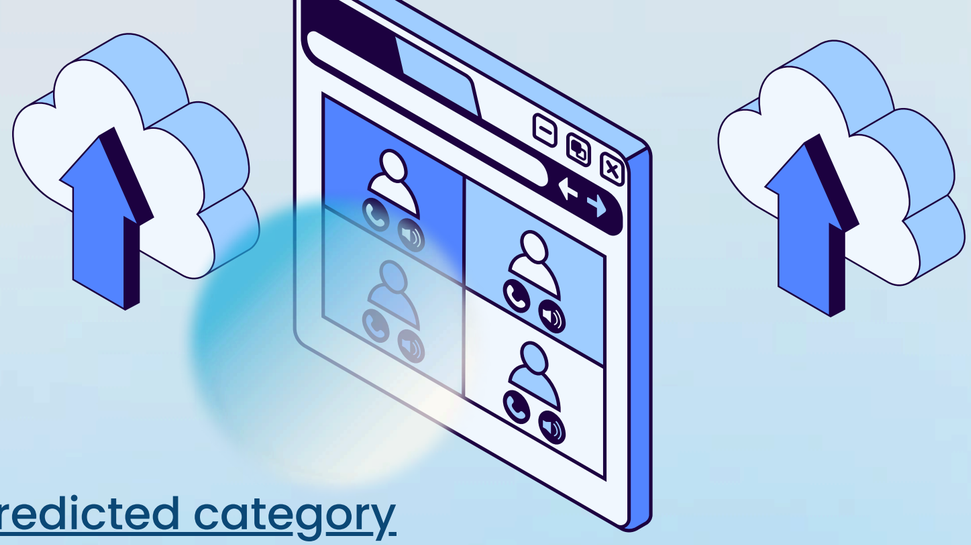
Macro average for F1-score : 66%

Macro average for recall : 67%

Macro average for precision : 71%

Natural Language Processing

Implementation



5.Second approach : Predicting the subcategory based on the predicted category

Step 1 :

Category classification

Step 2 :

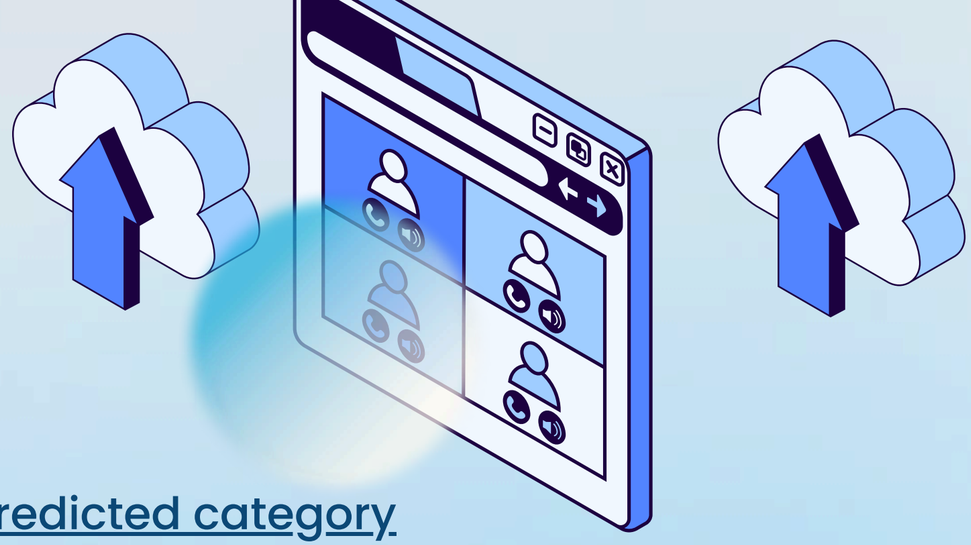
train Seperate classifiers for each category

Step 3 : Result

Prediction for Subcategories

Natural Language Processing

Implementation



5.Second approach : Predicting the subcategory based on the predicted category

SVM Model : (Linear Kernel)

	precision	recall	f1-score	support
NUM|dist	1.00	0.44	0.61	16
NUM|date	1.00	0.98	0.99	47
NUM|weight	1.00	0.50	0.67	4
NUM|other	0.71	0.42	0.53	12
NUM|speed	1.00	0.83	0.91	6
NUM|temp	1.00	0.60	0.75	5
NUM|period	0.78	0.88	0.82	8
NUM|count	0.60	1.00	0.75	9
NUM|money	0.00	0.00	0.00	3
NUM|perc	1.00	0.67	0.80	3
LOC|city	1.00	0.78	0.88	18
LOC|other	0.88	0.76	0.82	50
LOC|mount	1.00	0.67	0.80	3
LOC|state	0.56	0.71	0.62	7
LOC|country	1.00	1.00	1.00	3
HUM|desc	0.00	0.00	0.00	3
HUM|gr	0.50	0.67	0.57	6
HUM|ind	0.80	0.95	0.87	55
HUM|title	0.00	0.00	0.00	1
DESC|desc	0.27	0.43	0.33	7
DESC|def	0.87	0.95	0.91	123
DESC|reason	1.00	0.67	0.80	6
...				
micro avg	0.76	0.75	0.76	500
macro avg	0.66	0.60	0.59	500
weighted avg	0.79	0.75	0.75	500

Accuracy : 75%

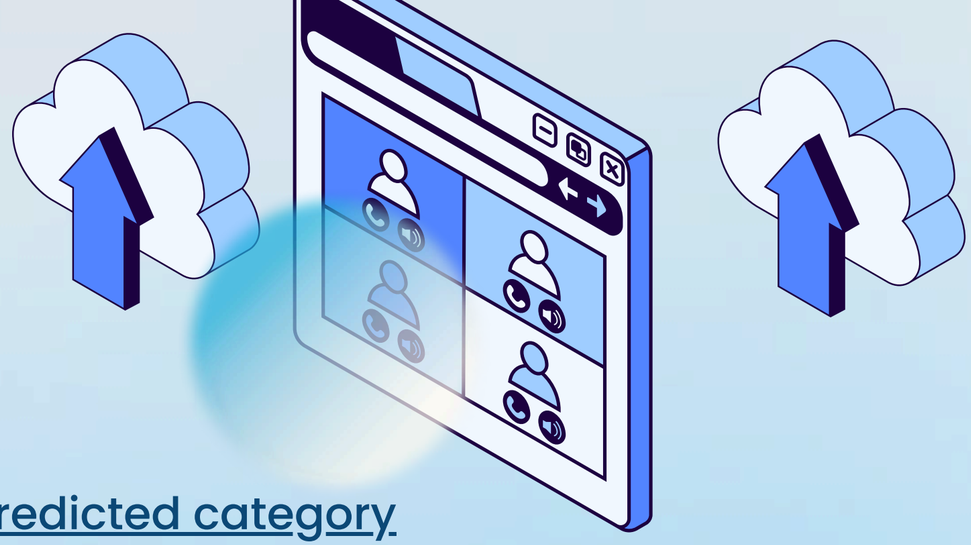
Macro average for recall : 60%

Macro average for F1-score : 59%

Macro average for precision : 66%

Natural Language Processing

Implementation



5.Second approach : Predicting the subcategory based on the predicted category

Combined Models :

	precision	recall	f1-score	support
NUM|dist	1.00	0.44	0.61	16
NUM|date	0.94	0.94	0.94	47
NUM|weight	1.00	0.50	0.67	4
NUM|other	0.83	0.42	0.56	12
NUM|speed	1.00	0.50	0.67	6
NUM|temp	1.00	0.80	0.89	5
NUM|period	0.64	0.88	0.74	8
NUM|count	0.64	1.00	0.78	9
NUM|money	0.50	0.33	0.40	3
NUM|perc	1.00	0.33	0.50	3
LOC|city	0.93	0.78	0.85	18
LOC|other	0.86	0.76	0.81	50
LOC|mount	1.00	0.67	0.80	3

LOC|state	0.40	0.57	0.47	7
LOC|country	1.00	1.00	1.00	3
HUM|desc	0.00	0.00	0.00	3
HUM|gr	1.00	0.50	0.67	6
HUM|ind	0.85	0.95	0.90	55
HUM|title	0.00	0.00	0.00	1
DESC|desc	0.75	0.86	0.80	7
DESC|def	0.78	0.98	0.87	123
DESC|reason	1.00	0.67	0.80	6
...				
micro avg	0.78	0.77	0.77	500
macro avg	0.75	0.60	0.62	500
weighted avg	0.82	0.77	0.76	500

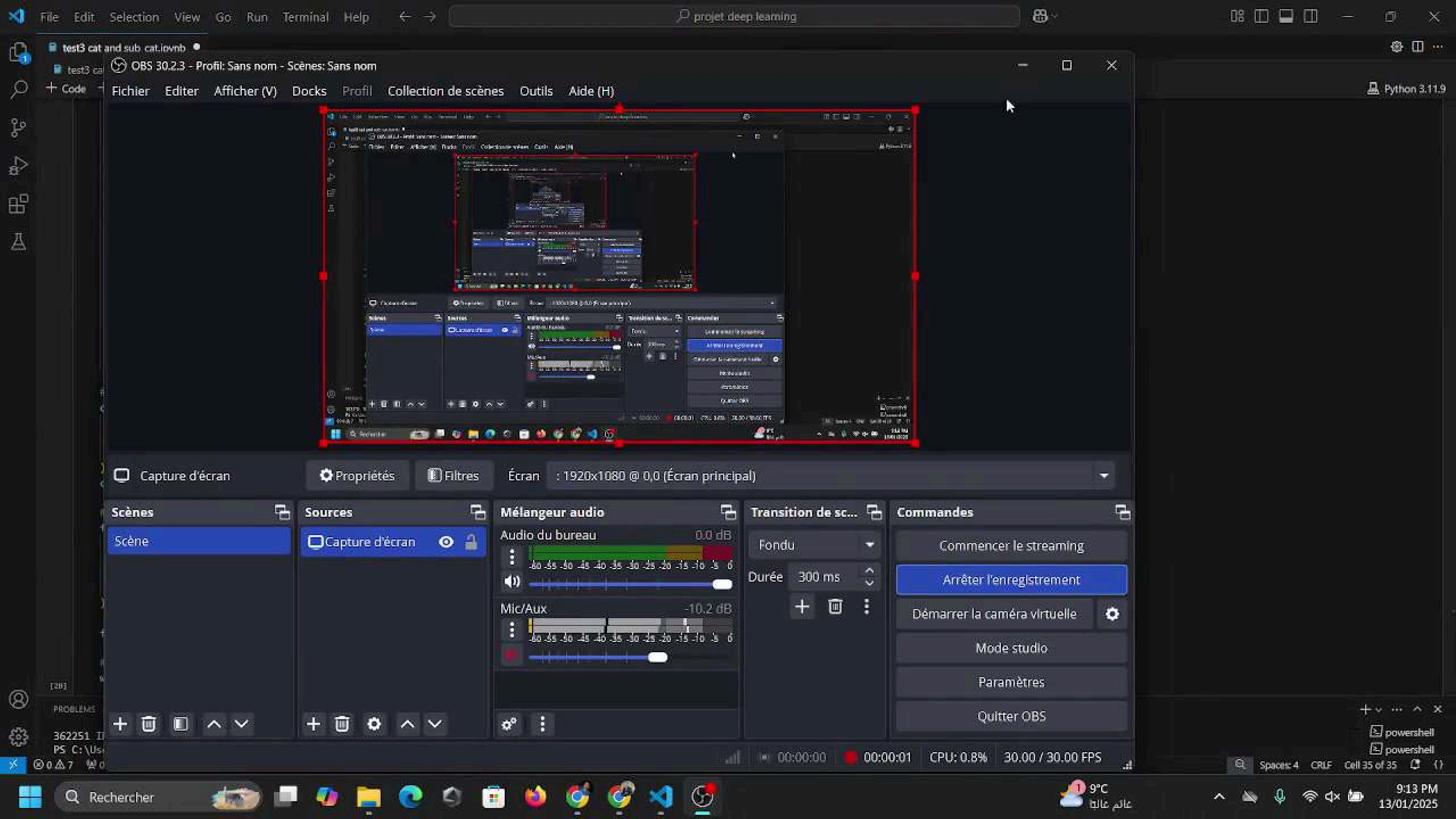
Accuracy : 77%

Macro average for recall : 60%

Macro average for F1-score : 62%

Macro average for precision : 75%

Video Demonstration


File Edit Selection View Go Run Terminal Help
projet deep learning
test3 cat and sub cat.ipynb
OBS 30.2.3 - Profil: Sans nom - Scènes: Sans nom
Fichier Editer Afficher (V) Docks Profil Collection de scènes Outils Aide (H)
Python 3.11.9
Capture d'écran
Propriétés
Filtres
Écran : 1920x1080 @ 0,0 (Écran principal)
Scènes
Scène
Sources
Capture d'écran
Mélangeur audio
Audio du bureau 0.0 dB
Mic/Aux -10.2 dB
Transition de sc...
Fondu
Durée 300 ms
Commandes
Commencer le streaming
Arrêter l'enregistrement
Démarrer la caméra virtuelle
Mode studio
Paramètres
Quitter OBS
00:00:00
00:00:01
CPU: 0.8%
30.00 / 30.00 FPS
powershell
powershell
Rechercher
9:13 PM
13/01/2025

Conclusion

	First Method					Second Method	
	SVM	Random Forest	Gradient Boosting	Combined Models	ANN	SVM	Combine Models
Accuracy	78%	75%	75%	77%	81%	75%	77%
F1 score	64%	53%	58%	62%	66%	59%	62%
Recall	64%	53%	58%	60%	67%	60%	60%
Precision	74%	63%	65%	76%	71%	66%	75%

Perspectives

Using pre-trained models like BERT, RoBERTa, or T5

Data Augmentation using generative models like GPT

Thank You for your Attention!