Republic of Tunisia
Ministry of Higher Education and Scientific Research
University of Tunis El Manar
National School of Engineers of Tunis

Deep Learning Project

From Classical to Cutting-Edge: Exploring Different Approcahes for Questions Classification

Realised by :
Trigui Hatem
Hassouna Malek
Med Rabii Baccari

Supervised by :
Ms. Linda Marrakchi
M. Mustapha BENHAJ MINIAOUI

Class : 3ATEL.DASEC

PLAN

01 General Overview & Preliminary operations

02 Use classical Machine Learning Approaches

03 Use Deep Learning Approaches

04 Use Transformer-Based Approach

05 Conclusion

General Overview & Preliminary operations

Dataset - TREC

Train Set

5452 Rows

train_set5.txt

```
DESC:manner How did serfdom develop in and then leave Russia ?
ENTY:cremat What films featured the character Popeye Doyle ?
DESC:manner How can I find a list of celebrities ' real names ?
ENTY:animal What fowl grabs the spotlight after the Chinese Year of the Monkey ?
```

Label

Question

Test Set

500 Rows

test_set.txt

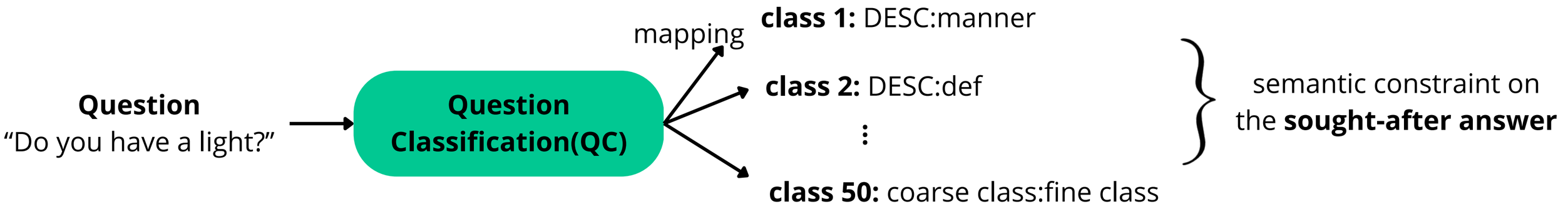
```
NUM:dist How far is it from Denver to Aspen ?
LOC:city What county is Modesto , California in ?
HUM:desc Who was Galileo ?
DESC:def What is an atom ?
NUM:date When did Hawaii become a state ?
```

6 coarse classes and **50 fine classes.**

Class	#	Class	#
ABBREV.	9	description	7
abb	1	manner	2
exp	8	reason	6
ENTITY	94	**HUMAN**	65
animal	16	group	6
body	2	individual	55
color	10	title	1
creative	0	description	3
currency	6	**LOCATION**	81
dis.med.	2	city	18
event	2	country	3
food	4	mountain	3
instrument	1	other	50
lang	2	state	7
letter	0	**NUMERIC**	113
other	12	code	0
plant	5	count	9
product	4	date	47
religion	0	distance	16
sport	1	money	5
substance	15	order	0
symbol	0	other	12
technique	1	period	8
term	7	percent	5
vehicle	4	speed	6
word	0	temp	5
DESCRIPTION	138	size	0
definition	123	weight	4

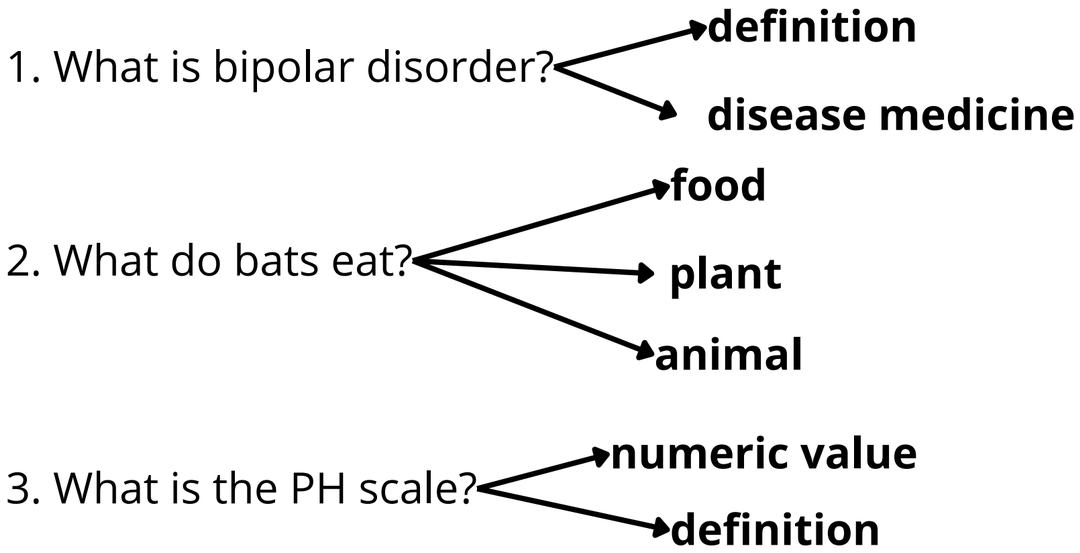
The distribution of 500 TREC 10 (**test set**) questions over the question hierarchy. Coarse classes (in bold) are followed by their fine class refinements.

Purpose



The intension is that this we **do not** address questions like, **which calls for an action**, but rather **only factual Wh-questions**.

Challenges : The Ambiguity Problem



Multi-Label Classification

we allow our classifiers to assign multiple class labels for a single question.

Pre-processing

Question

How can I find a list of celebrities ' real names ?

1. **lower case**

how can I find a list of celebrities ' real names ?

Use **regex**

2. Remove non-alphabetic characters

how can I find a list of celebrities real names

Use **word_tokenize** from nltk.tokenize

3. Tokenize

['how', 'can', 'I', 'find', 'a', 'list', 'of', 'celebrities', 'real', 'names']

Use **stopwords** from nltk.corpus **- {"what", "where", "who", "when", "why", "how"}**

4. remove stop words excluding wh questions

['how', 'find', 'list', 'celebrities', 'real', 'names']

Use **WordNetLemmatizer** from nltk.stem

5. Lemmatization

['how', 'find', 'list', '**celebrity**, 'real', '**name**']

Join

how find list celebrity real name

Embedding

Label

DESC:manner

1. **Label Splitting**

category
DESC

specific_type
manner

more granularity

Use **LabelEncoder** from sklearn.preprocessing

2. Encoding both categories and specific types

Use **LabelEncoder** from sklearn.preprocessing

3. Encoding combined label

category	specific_type	category_encoded	specific_type_encoded
DESC	manner	1	23

combined_label	combined_label_encoded
DESC_manner	4

Train Set Overview (after preprocessing + encoding)

	question	label	processed_question	category	specific_type	category_encoded	specific_type_encoded	combined_label	combined_label_encoded
0	How did serfdom develop in and then leave Russ...	DESC:manner	how serfdom develop leave russia	DESC	manner	1	23	DESC_manner	4
1	What films featured the character Popeye Doyle ?	ENTY:cremat	what film featured character popeye doyle	ENTY	cremat	2	8	ENTY_cremat	9
2	What fowl grabs the spotlight after the Chines...	ENTY:animal	what fowl grab spotlight chinese year monkey	ENTY	animal	2	1	ENTY_animal	6

Use classical Machine Learning Approaches

I . First Word Representation Approach (TF-IDF)

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

General Process

train_set['processed_question']
test_set['processed_question']

Labels:
train_set['combined_label_encoded']

1. Feature Extraction: TF-IDF Vectorization

Features: X_train: TF-IDF matrix from training set

Parameters
max_features=5000: Limits the number of features (words) to 5,000 to reduce dimensionality and avoid overfitting.

2.Model Training

trained model

3. Prediction

Features: X_test (TF-IDF matrix from the test set).

y_pred: Predicted labels

4. Evaluation

True labels:
test_set['combined_label_encoded']

I . First Word Representation Approach (TF-IDF)

Classification Report

50 classes

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	1.00	0.62	0.77	8
2	0.74	0.98	0.85	123
49	0.00	0.00	0.00	4
Accuracy			**0.71**	500
macro avg	0.53	0.45	0.45	500
weighted avg	**0.70**	**0.71**	**0.66**	**500**

Since our data is **imbalanced** so we'll consider weighted avg values weighing each class's contribution by its support (i.e., the number of true instances in that class)

Logistic Regression model

I . First Word Representation Approach (TF-IDF)

Imbalanced Data ?

category

ENTY 1250
HUM 1223
DESC 1162
NUM 896
LOC 835
ABBR 86

dominate the dataset

minority class

specific_type

ind 962
other 733
def 421
count 363
desc 321
manner 276
date 218
...
letter 9
code 9
speed 9
temp 8
ord 6
religion 4
currency 4

most common types

minority classes

1. Class Imbalance:

Minority classes may lead to poor model performance for those classes.

2. Granularity:

Some specific types, such as techmeth and volsize, are too granular, which may increase the complexity of classification.

First Solution (Oversampling)

- X_train: The feature matrix of the training set (inputs for the model).
- y_train: train_set['combined_label_encoded']

smote object

1. Initializing the SMOTE Object

k_neighbors=3: NNN to use when generating synthetic samples for the minority class.

2.Applying SMOTE to the Training Data

For each minority class sample, 3 nearest neighbors will be found, and synthetic samples will be created based on these neighbors.

X_train_balanced
y_train_balanced

equal number of samples for each class (or at least closer to equal)

I . First Word Representation Approach (TF-IDF)

First Solution (Oversampling)

Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	0.67	0.75	0.71	8
2	0.88	0.76	0.82	123
49	1.00	1.00	1.00	4
Accuracy			**0.76**	500
macro avg	0.68	0.72	0.67	500
weighted avg	**0.82**	**0.76**	**0.77**	**500**

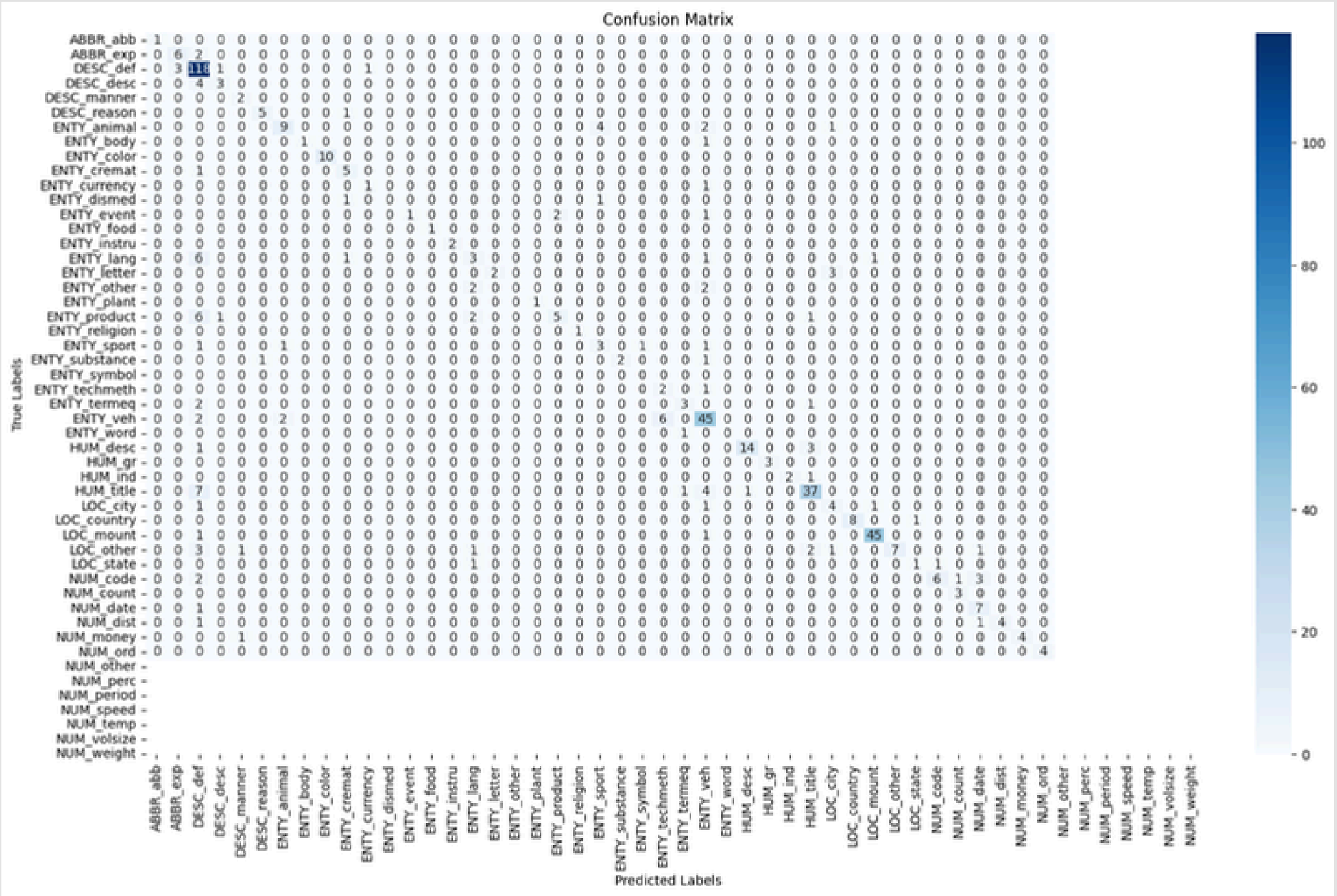
Logistic Regression model

Classification Report

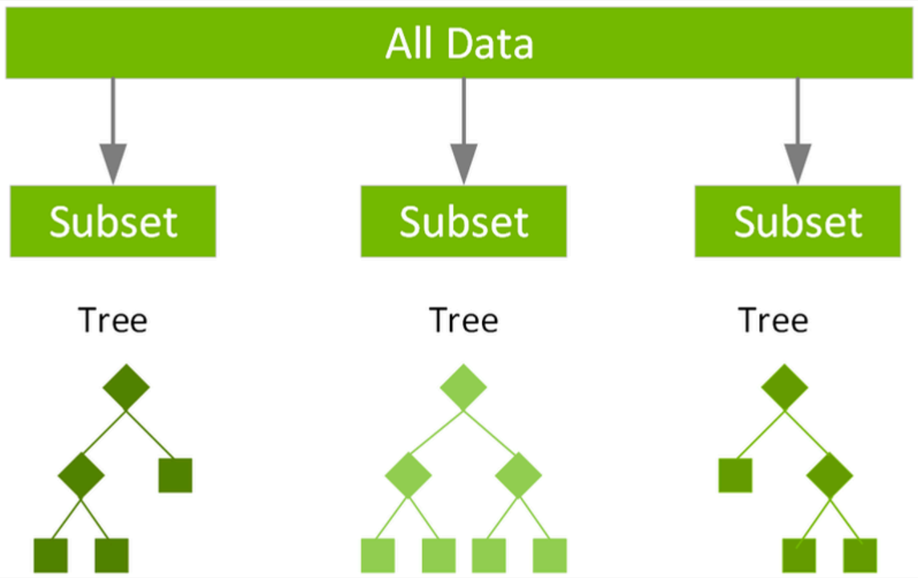
	precision	recall	f1-score	support
0	1.00	1.00	1.00	1
1	0.67	0.75	0.71	8
2	0.74	0.96	0.84	123
49	1.00	1.00	1.00	4
Accuracy			**0.76**	500
macro avg	0.72	0.64	0.66	500
weighted avg	**0.77**	**0.76**	**0.75**	**500**

Random Forest model

I . First Word Representation Approach (TF-IDF)



Second Solution (Use Imbalanced data supportive model “XGboost”)



- XGBoost assigns higher weights to the minority class using the **scale_pos_weight** parameter.

- Decision trees split the data to maximize information gain, and XGBoost focuses on reducing errors iteratively.

Classification Report

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.75	0.75	0.75	8
2	0.75	0.99	0.86	123
49	1.00	0.75	0.86	4
Accuracy			**0.76**	500
macro avg	0.66	0.59	0.59	500
weighted avg	**0.78**	**0.77**	0.75	**500**

XGBoost model

I . First Word Representation Approach (TF-IDF)

XGBoost's hyperparameters tuning

1. **Define Parameter Grid**
- **max_depth:** [3, 6, 10]
- **learning_rate:** [0.01, 0.1, 0.3]
- **n_estimators:** [100, 200, 500]

2. **Perform Grid Search** (cv=3)
3. **Select the Best Model**

Best Parameters: {'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 200}

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.75	0.75	0.75	8
2	0.73	0.99	0.86	123
49	1.00	0.75	0.86	4
Accuracy			**0.76**	500
macro avg	0.66	0.57	0.86	500
weighted avg	**0.78**	**0.76**	**0.75**	**500**

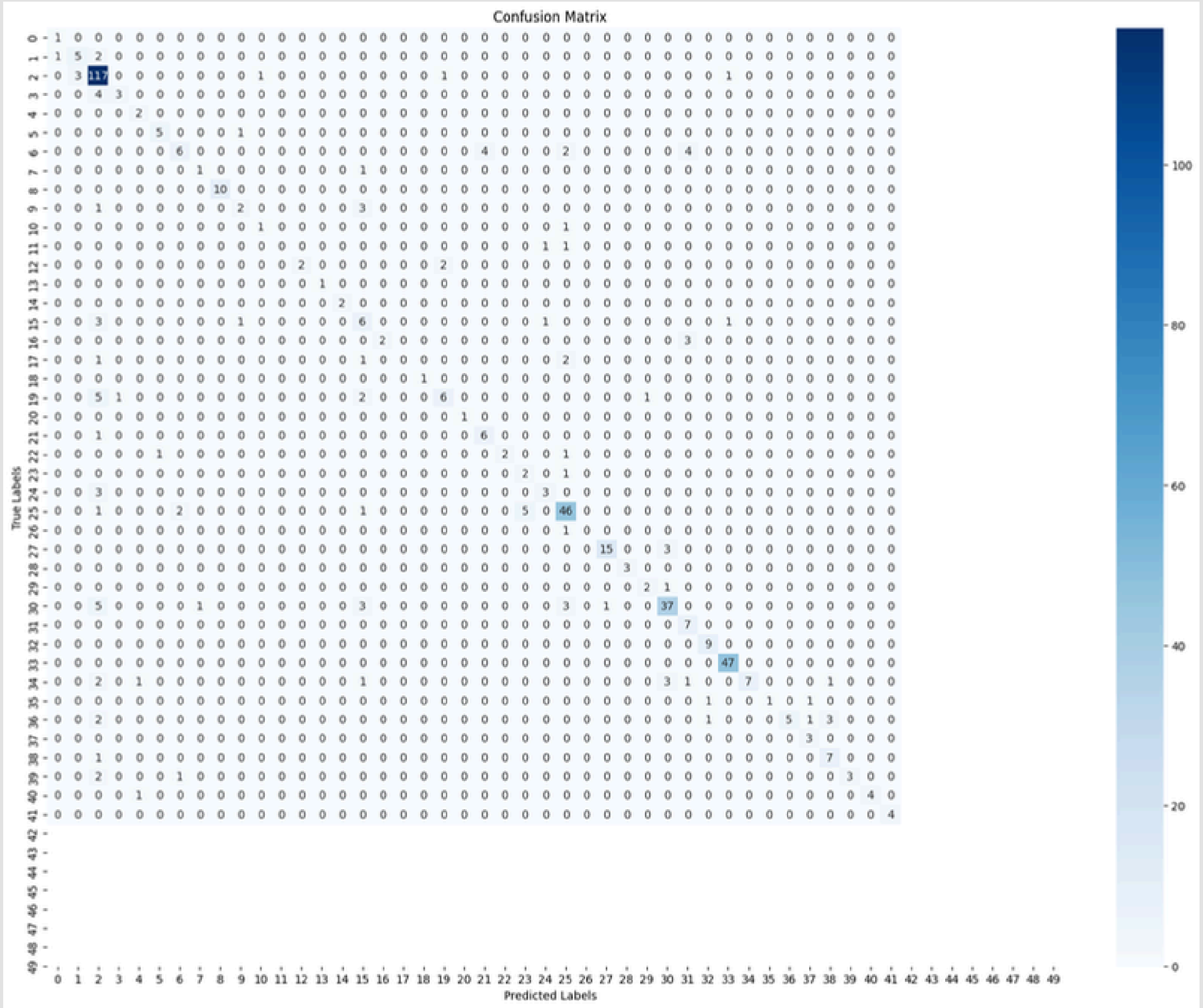
SMOTE + XGBoost

XGBoost works well when **combined** with SMOTE to balance the dataset before training to improve the **representation of the minority classes** while still benefiting from XGBoost's **efficient learning process.**

	precision	recall	f1-score	support
0	0.50	1.00	0.67	1
1	0.62	0.62	0.62	8
2	0.78	0.95	0.86	123
49	1.00	1.00	1.00	4
Accuracy			**0.77**	500
macro avg	0.72	0.67	0.67	500
weighted avg	**0.79**	**0.77**	**0.76**	**500**

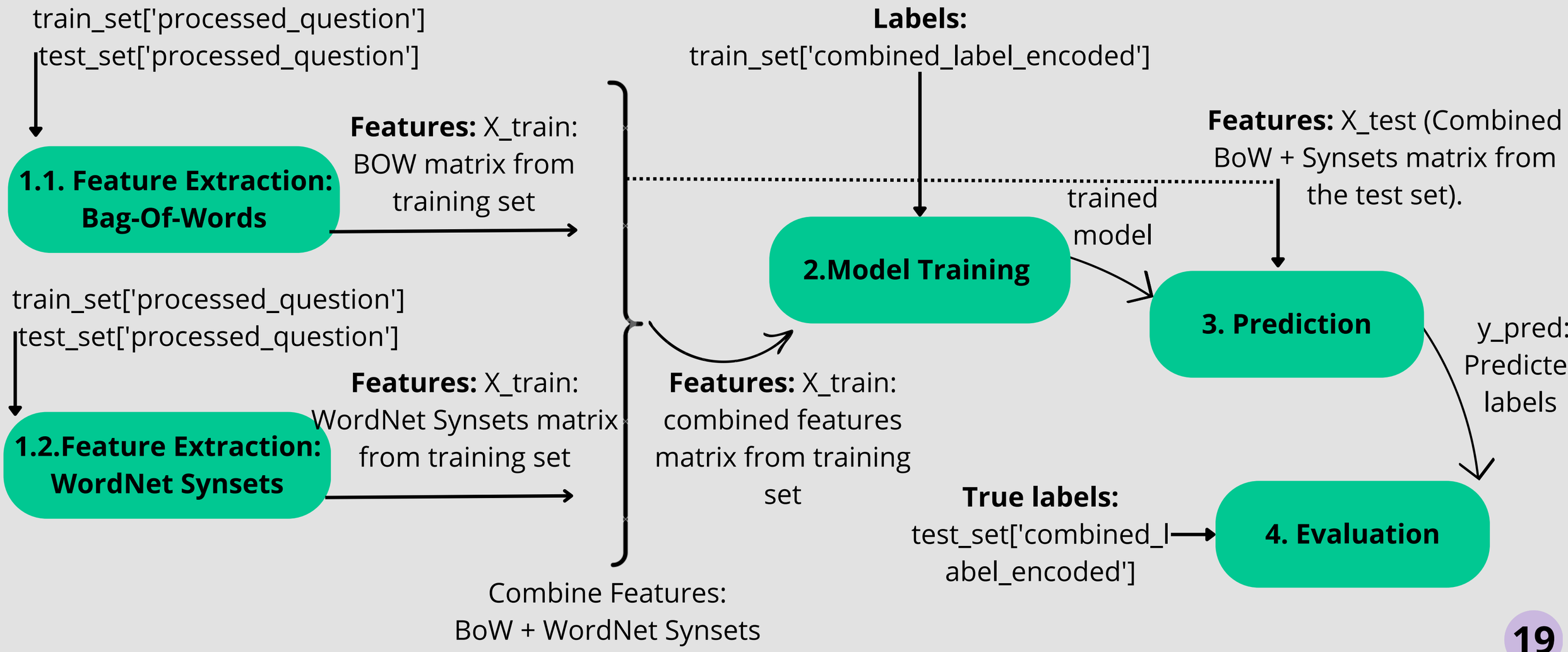
I . First Word Representation Approach (TF-IDF)

SMOTE + XGBoost



II. Second Word Representation Approach (BOW + WordNet Synsets):

- **BoW focuses on word frequency but misses context.**
- **WordNet captures semantic context but might generalize too much.**



I . First Word Representation Approach (TF-IDF)

Classification Report

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.86	0.75	0.80	8
2	0.77	1.00	0.87	123
49	1.00	0.75	0.86	4
Accuracy			**0.81**	500
macro avg	0.75	0.65	0.66	500
weighted avg	**0.83**	**0.81**	**0.80**	**500**

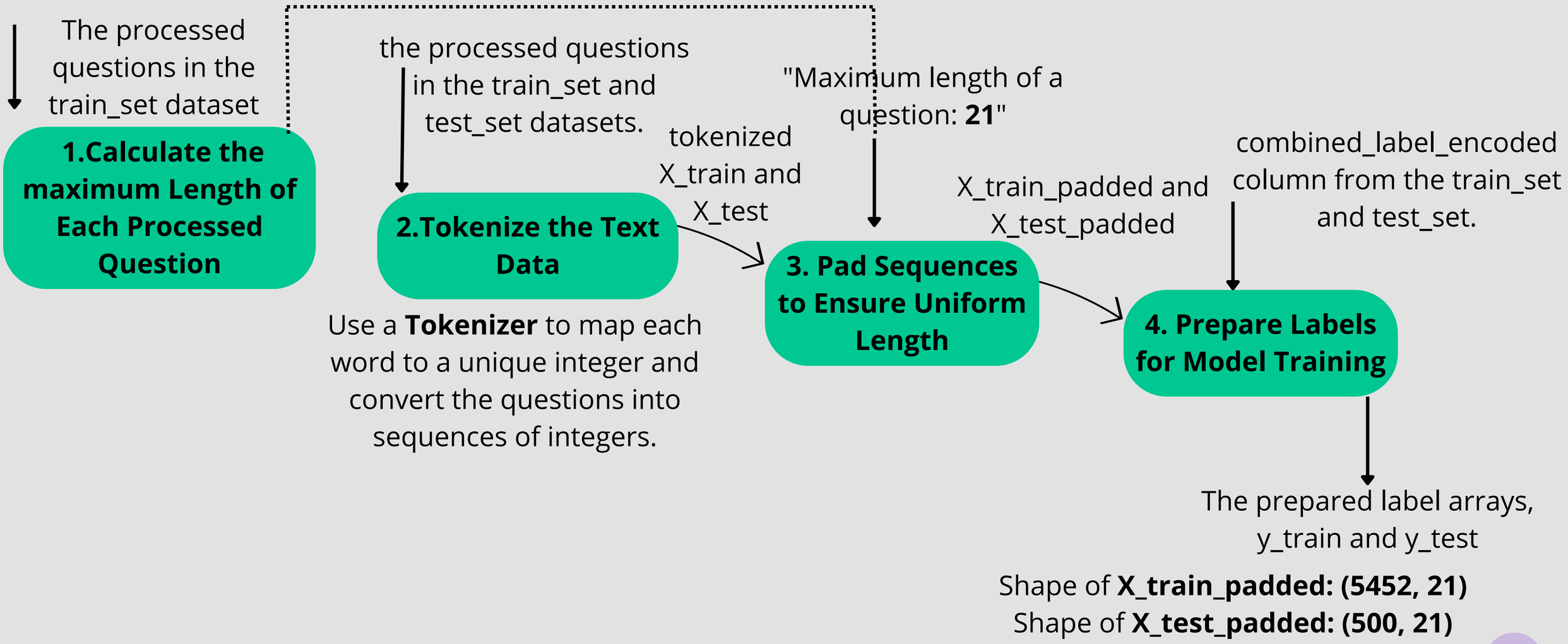
Support Vector Machine: SVM model

Recap

Approach	Embedding Method (Feature Extraction)	Classifier/model	precision	recall	f1-score	Accuracy
Classical Machine Learning	TF-IDF	Logistic Regression	0.70	0.71	0.66	0.71
		Logistic Regression + SMOTE	0.82	0.76	0.77	0.76
		Random Forest + SMOTE	0.77	0.76	0.75	0.76
		XGBoost	0.78	0.77	0.75	0.76
		XGBoost + hyperparameters tuning	0.78	0.76	0.75	0.76
		XGBoost + SMOTE	0.79	0.77	0.76	0.77
	BOW + WordNet Synsets	SVM	0.83	0.81	0.80	0.81

Use Deep Learning Approaches

Pre-processing



I. First Approach (Embedding: Use Embedding Layer)

Initial model (baseline)

sequential data (text sequences)

Embedding Layer

Dense vectors of dimension 100 for each token.

Helps the model understand relationships between words in a continuous space.

LSTM Layer

Contains 128 units and processes the input sequence to extract meaningful patterns.

Dropout Layer

Dropout Rate: 50% of the neurons are randomly deactivated during training.

Dense Layers

1. Hidden Dense Layer: Contains 64 neurons with ReLU activation
2. Output Dense Layer: Contains 50 neurons (softmax)

I. First Approach (Embedding: Use Embedding Layer)

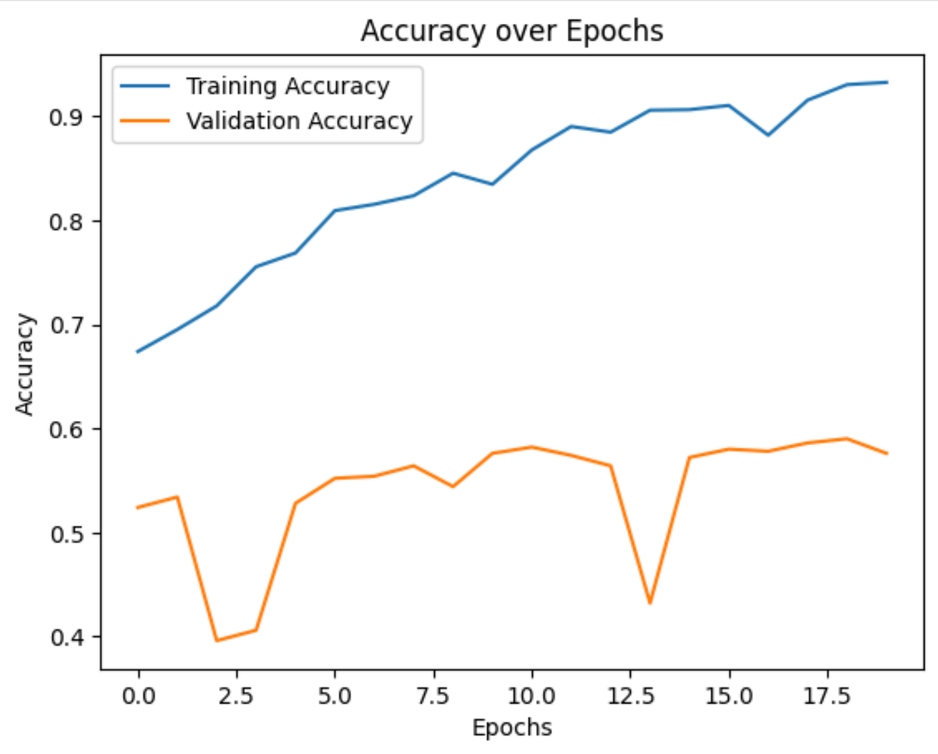
Initial model (baseline)

Test Accuracy: **57.60%**

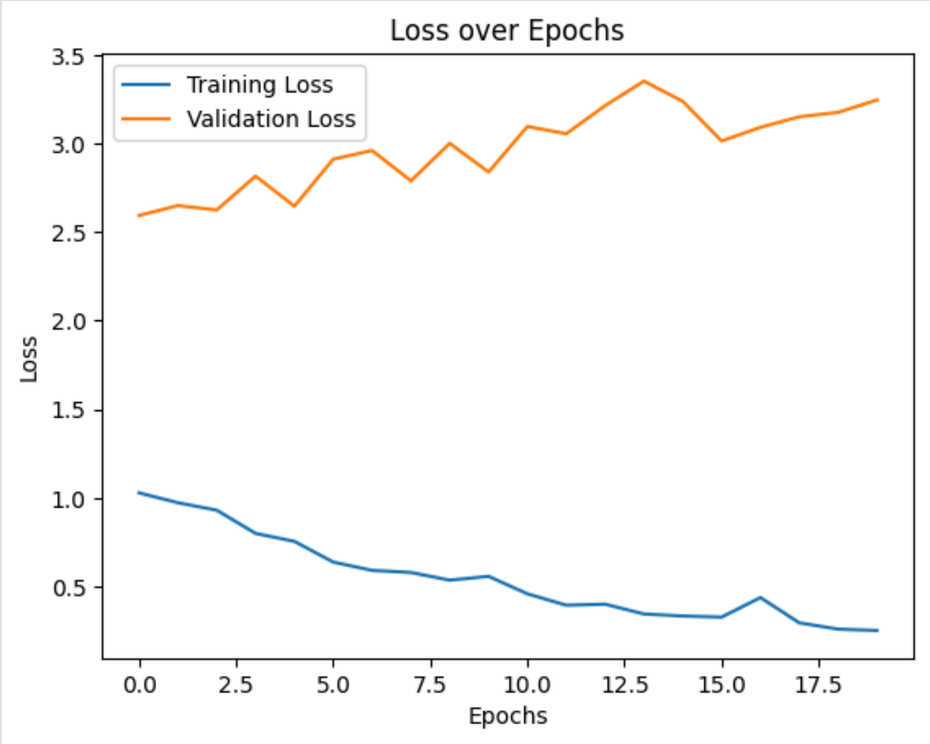
Question	Predicted label
How far is it from Denver to Aspen ?	ENTY_body
What county is Modesto , California in ?	LOC_other
Who was Galileo ?	HUM_desc
What is an atom ?	DESC_desc
When did Hawaii become a state ?	NUM_date
How tall is the Sears Building ?	DESC_manner
George Bush purchased a small interest in which baseball team **?**	DESC_def
What is Australia 's national flower ?	NUM_speed
Why does the moon turn orange ?	DESC_reason
What is autism ?	DESC_def

I. First Approach (Embedding: Use Embedding Layer)

Initial model (baseline)



high training accuracy

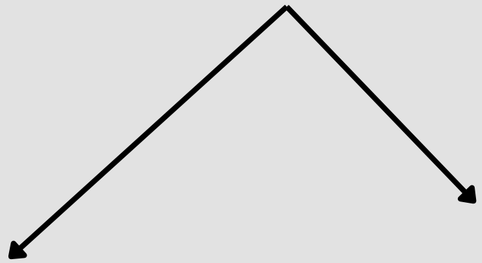


Overfitting

I. First Approach (Embedding: Use Embedding Layer)

Step 1 : Class Weights for Imbalanced Data

Used class weights to address class imbalance with class_weight from sklearn.utils to make it balanced



Still Overfitting

Test Accuracy: **48.80%**

Question	Predicted label
How far is it from Denver to Aspen ?	**NUM_dist**
What county is Modesto , California in ?	LOC_other
Who was Galileo ?	HUM_desc
What is an atom ?	**ABBR_exp**
When did Hawaii become a state ?	NUM_date
How tall is the Sears Building ?	**NUM_dist**
George Bush purchased a small interest in which baseball team **?**	**NUM_date**
What is Australia 's national flower ?	**ENTY_plant**
Why does the moon turn orange ?	DESC_reason
What is autism ?	**DESC_desc**

was wrong but wrongly classified it with another class

I. First Approach (Use Embedding Layer)

Step	Key changes	Results
Adding L2 Regularization to LSTM	LSTM Layer: Added L2 regularization (kernel_regularizer=l2(0.01))	**No significant improvement, overfitting still present.**
Increasing Dropout	Dropout Rate: Increased to 0.6 (Dropout(0.6)).	**Reduced overfitting slightly, but still not enough to generalize well to validation data.**
Using Adam Optimizer with Smaller Learning Rate:	Optimizer: Adam optimizer with a reduced learning rate (learning_rate=0.0005).	**Slower learning curve, but overfitting persisted.**
Adding L1 and L2 Regularization Together	LSTM Layer: Used both L1 and L2 regularization (kernel_regularizer=l1_l2(l1=0.01, l2=0.01)).	**Reduced some overfitting, but it remained evident.**
Tested Various Batch Sizes	Experimented with different batch sizes, including 64 and 32.	**Slight differences in performance, but no major effect on overfitting.**
Reduced LSTM Units (Smaller Model)	Reduced the number of LSTM units to 64 (down from 128).	**Slight improvement in generalization, but overfitting persisted.**
Learning Rate Schedulers	Learning Rate Scheduler: Attempted to adjust the learning rate dynamically using a scheduler.	**Slower convergence, but did not solve overfitting.**

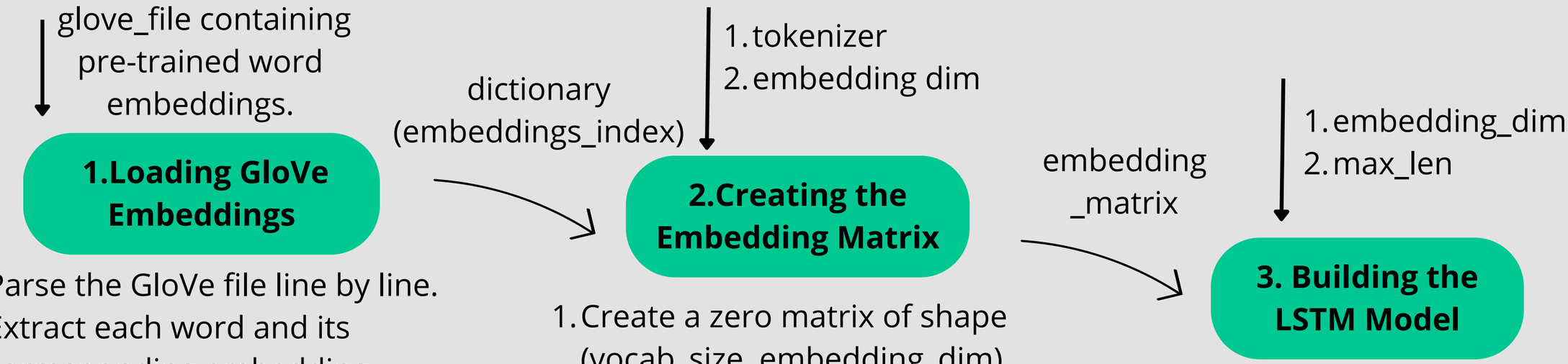
II. Second Approach (Use Pretrained GloVe Embeddings)

GloVe: Global Vectors for Word Representation : an **unsupervised learning** algorithm for obtaining vector representations for words. (deeper contextual embeddings)

Wikipedia 2014 + Gigaword 5 (**6B tokens**, **400K vocab**, uncased, **50d, 100d, 200d, & 300d vectors**)

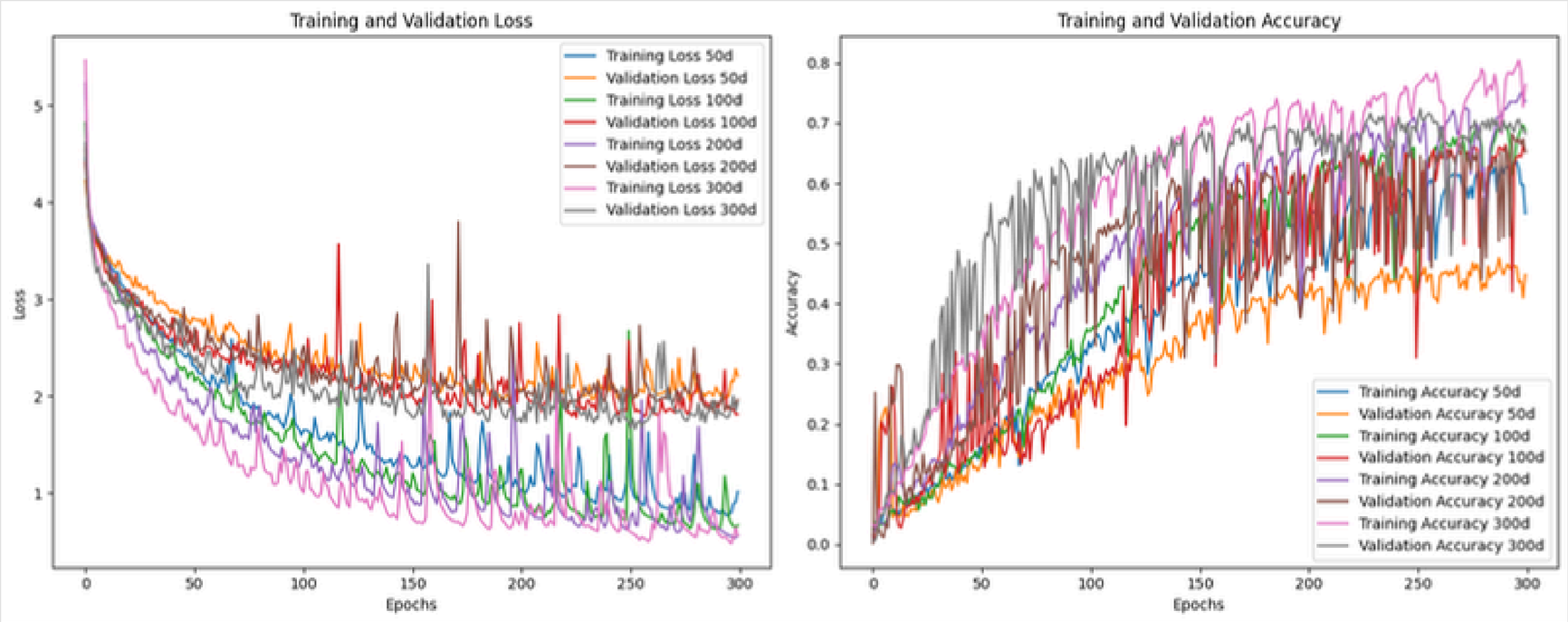
<word> <value_1> <value_2> ... <value_N>

<word>: The word or token.
<value_1>, <value_2>, ..., <value_N>: The numerical values representing the word's embedding in an N-dimensional space.



Embedding Layer: Uses the embedding_matrix and freezes its weights.

II. Second Approach (Use Pretrained GloVe Embeddings)

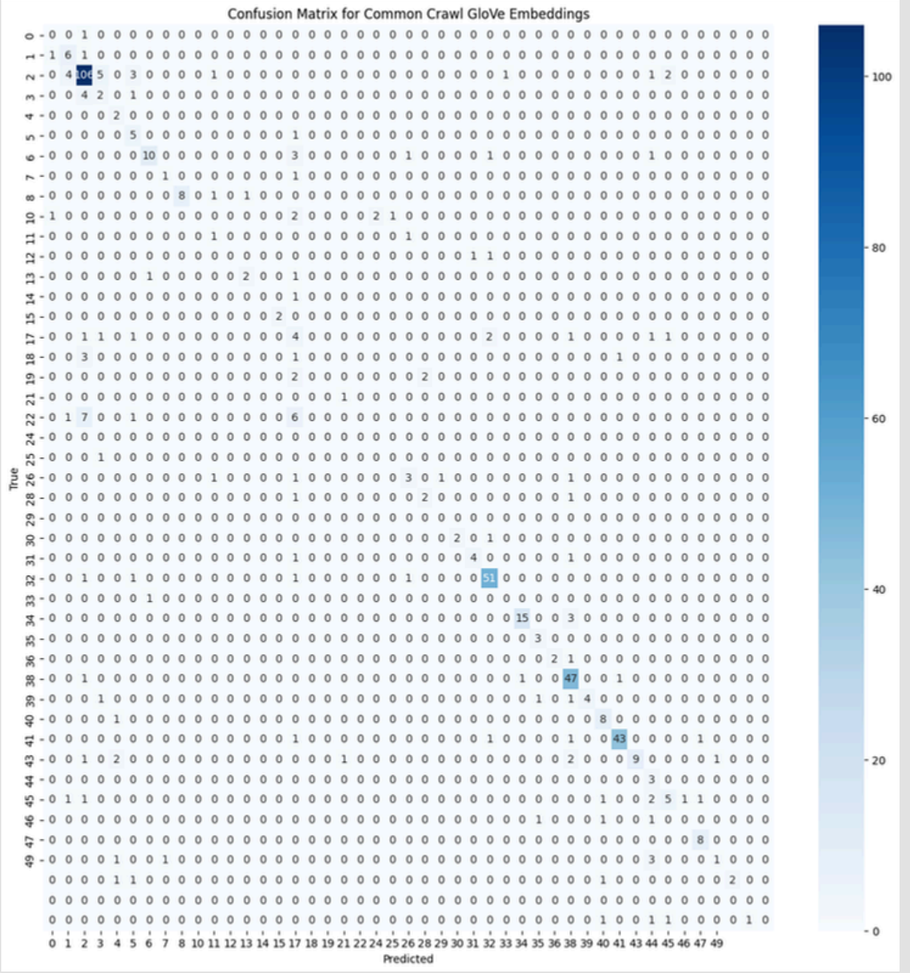


On **test Set**

	precision	**recall**	**f1-score**	**Accuracy**
50d GloVe	0.6785	0.4480	0.4621	0.4480
100d GloVe	0.7198	0.6520	0.6732	0.6520
200d GloVe	0.7455	0.6520	0.6792	0.6520
300d GloVe	0.7564	0.6820	0.6993	0.6820

II. Second Approach (Use Pretrained GloVe Embeddings)

Common Crawl (**840B tokens**, **2.2M vocab**, cased, **300d vectors**)



Evaluation Metrics:

Accuracy: 0.7180

F1 Score: 0.6973

Precision: 0.7011

Recall: 0.7180

Use Transformer-Based Approach

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

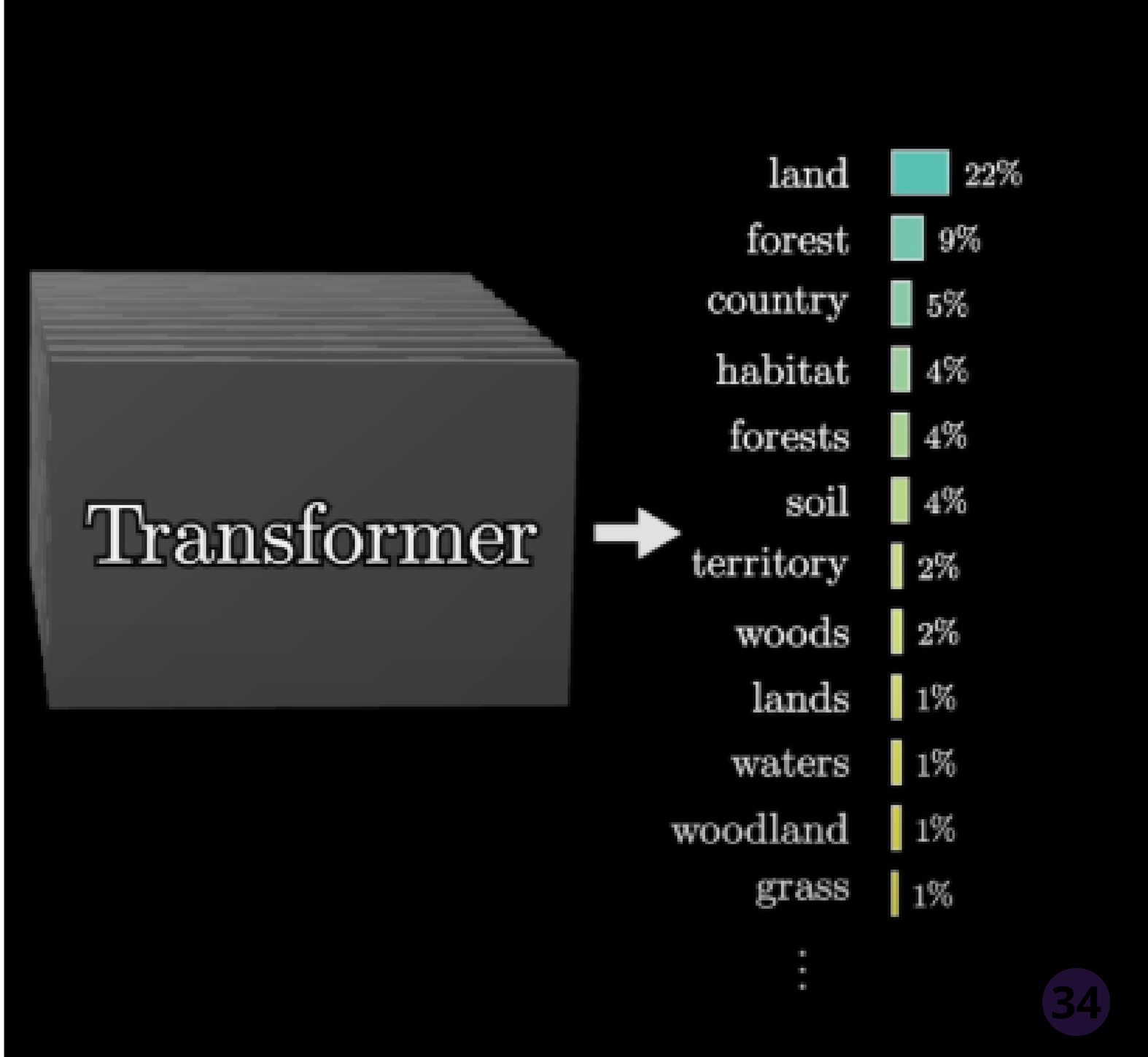
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to

Input Sequence:
Behold, this wild creature foraging on his _____ .



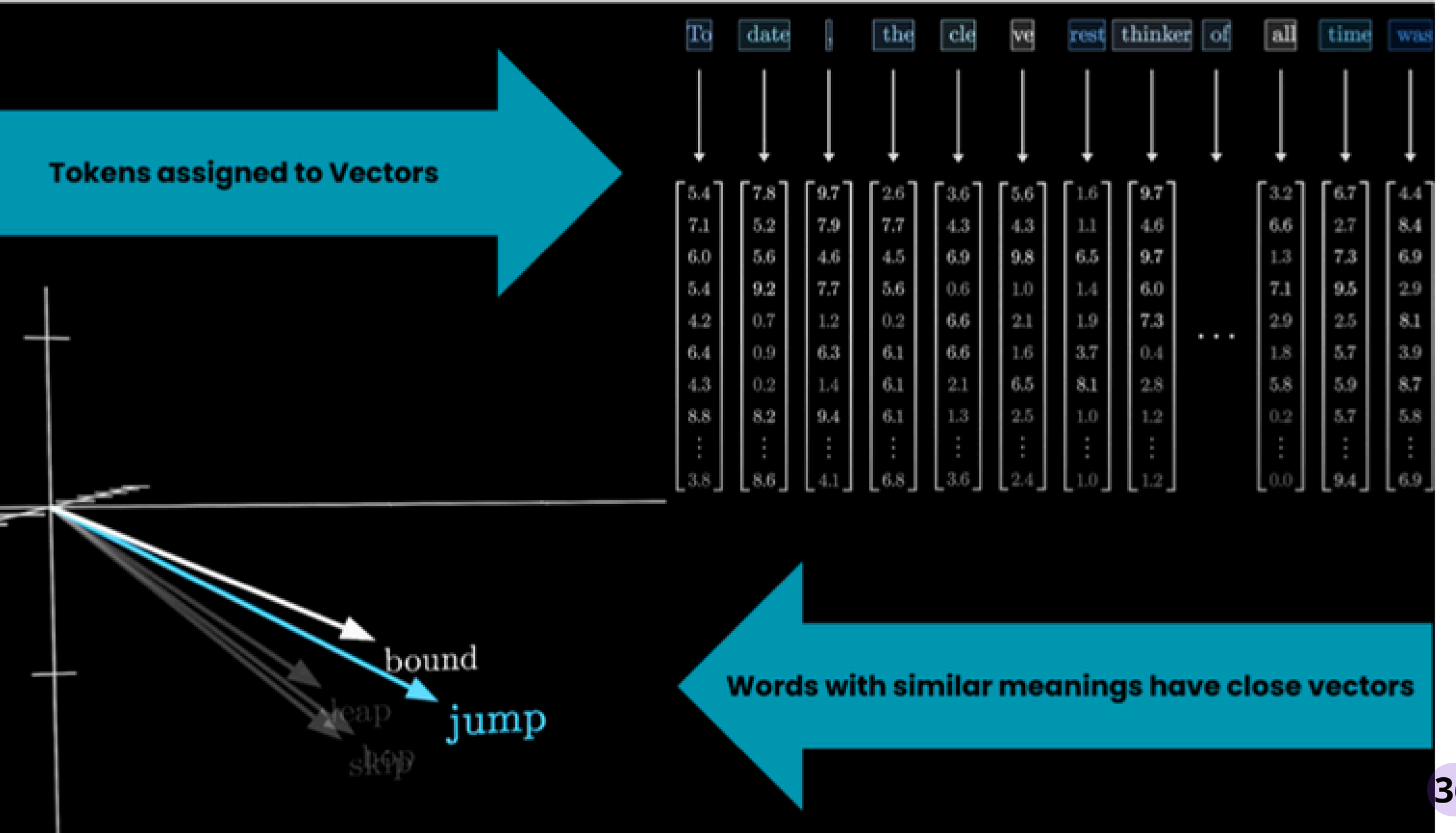
But First We tokenize :

To date, the cleverest thinker of all time was ...

The tokenization of this input would be:

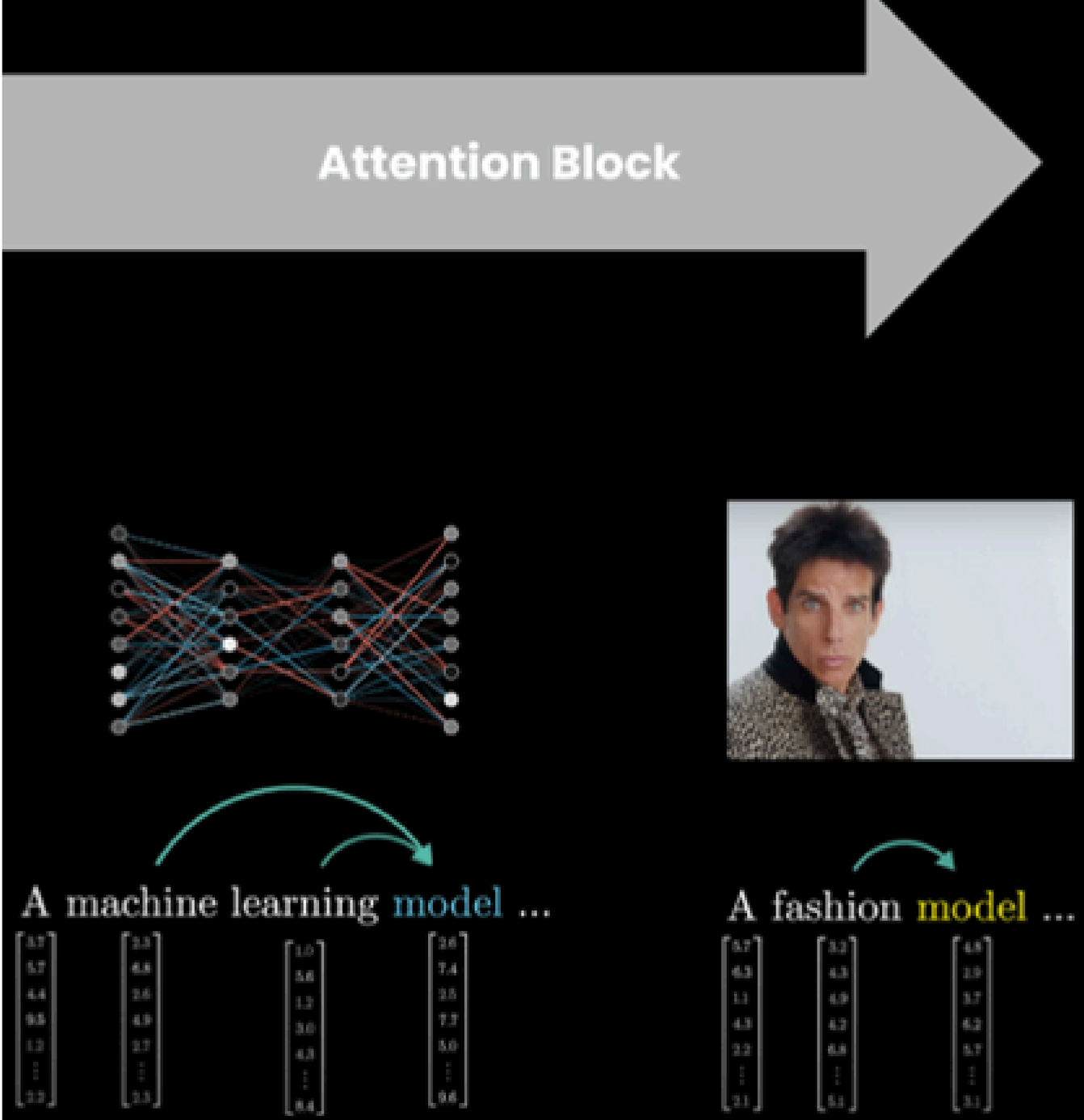
To| date|,| the| cle|ve|rest| thinker| of| all| time| was ...

Tokens assigned to Vectors



Words with similar meanings have close vectors

Attention Block

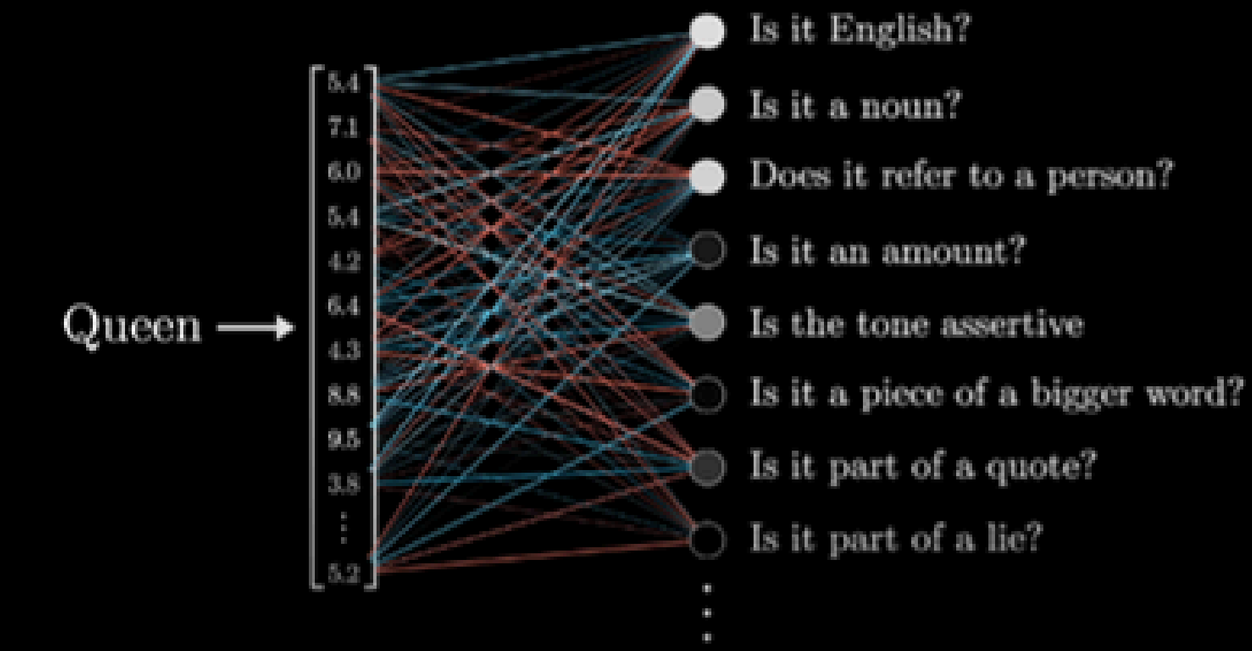


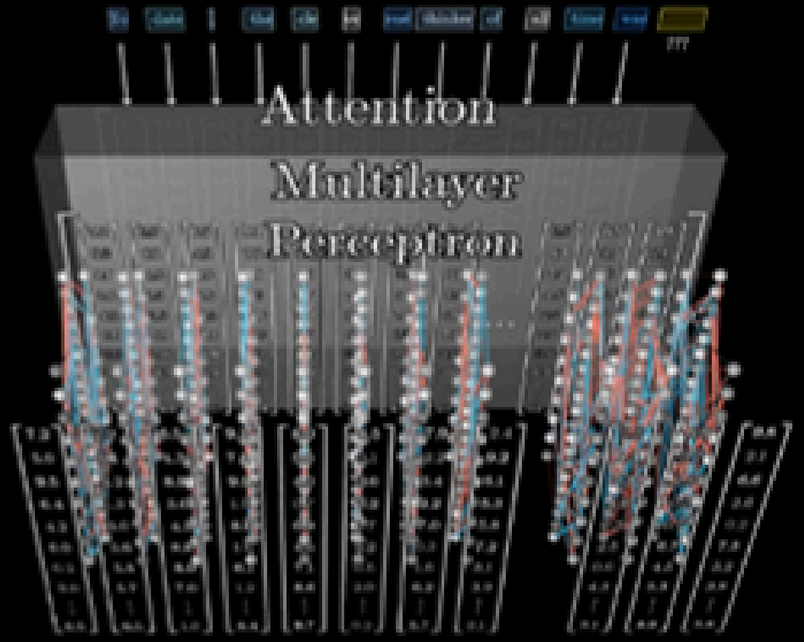
A machine learning model ...

A fashion model ...

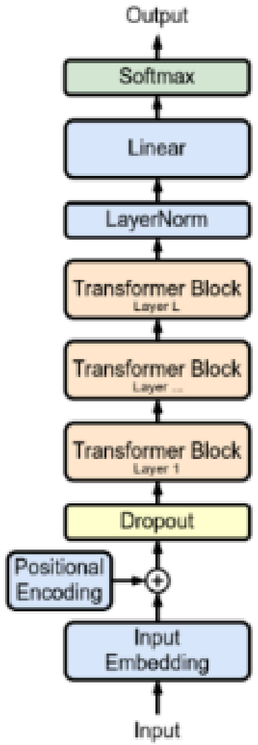
Example of Attention Block changing the vector of the word “model”

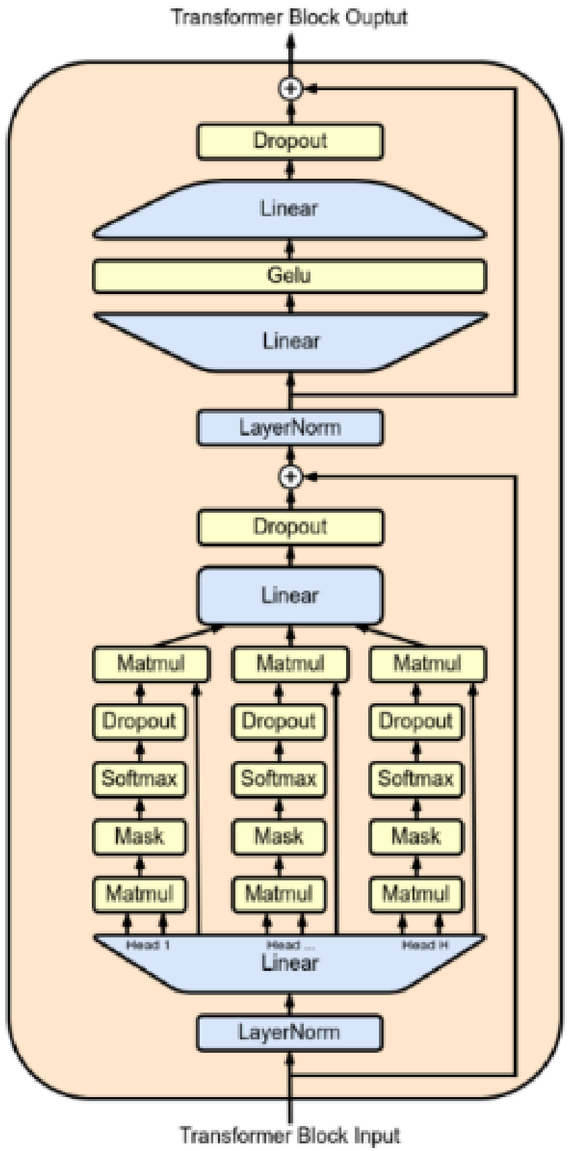
The Multilayer Perceptron asking "Questions" to Update the Vector

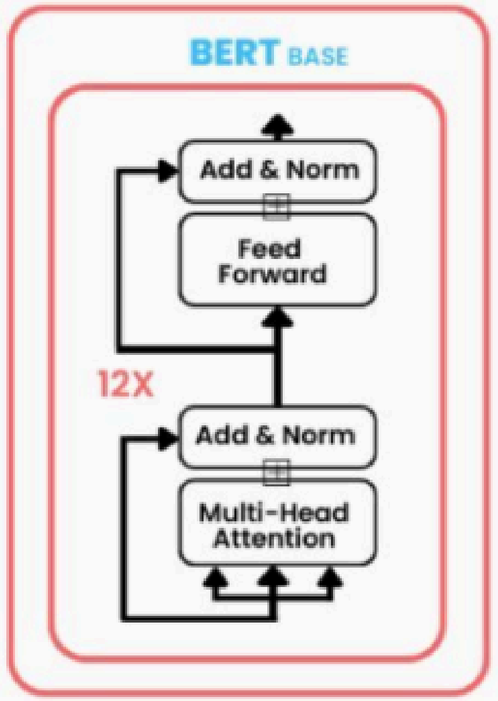




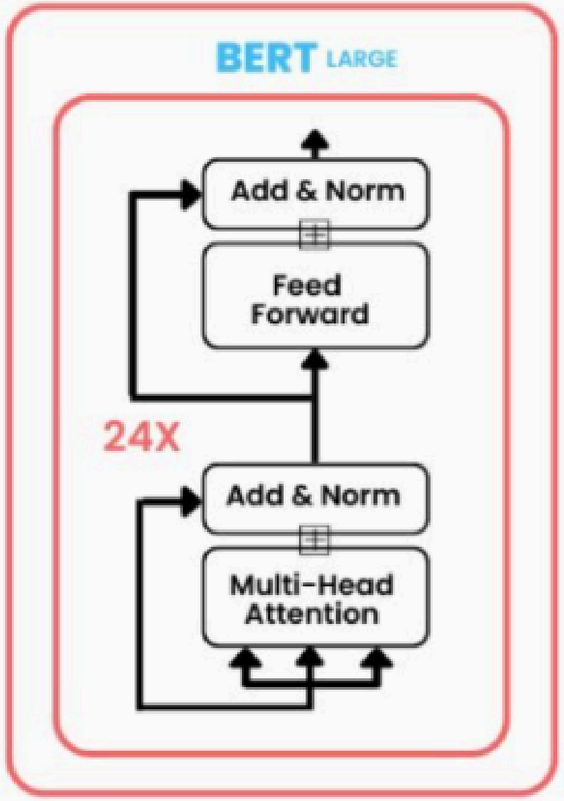
The Multilayer Perceptron "block"





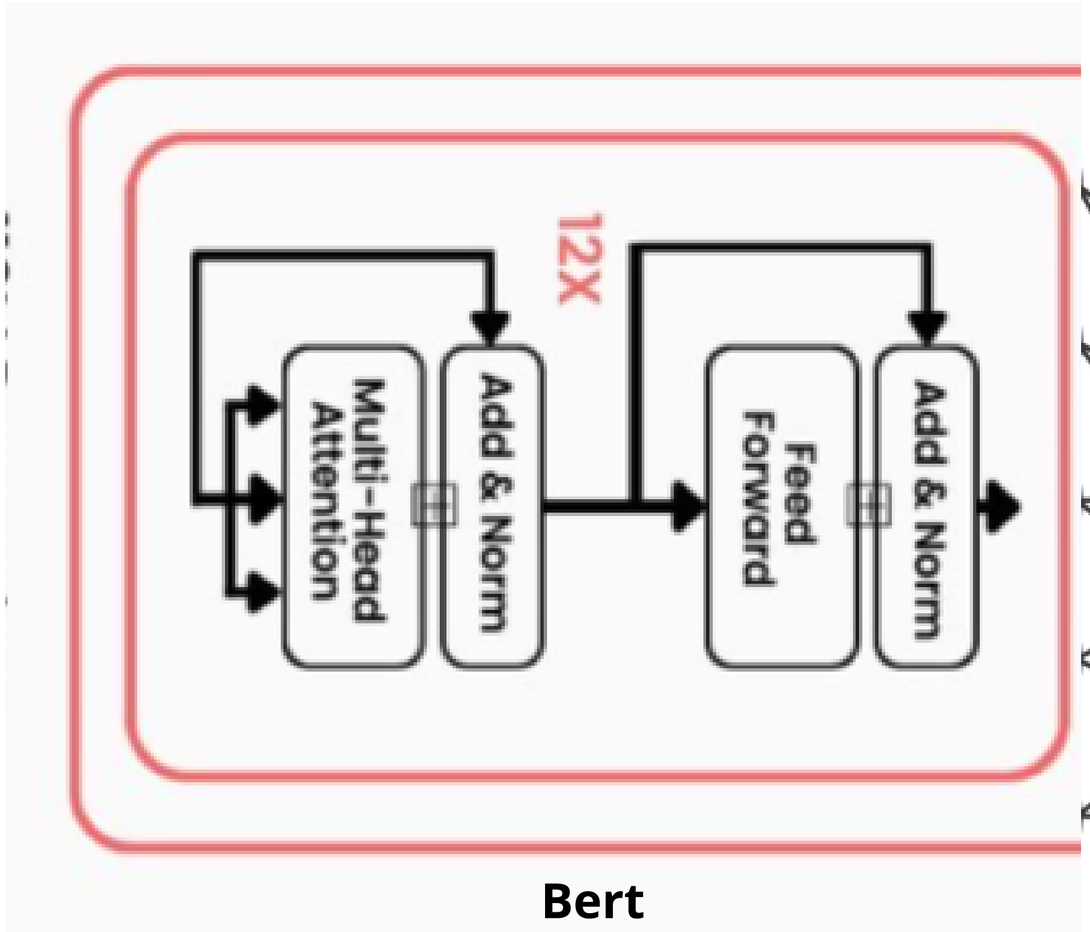


110M Parameters

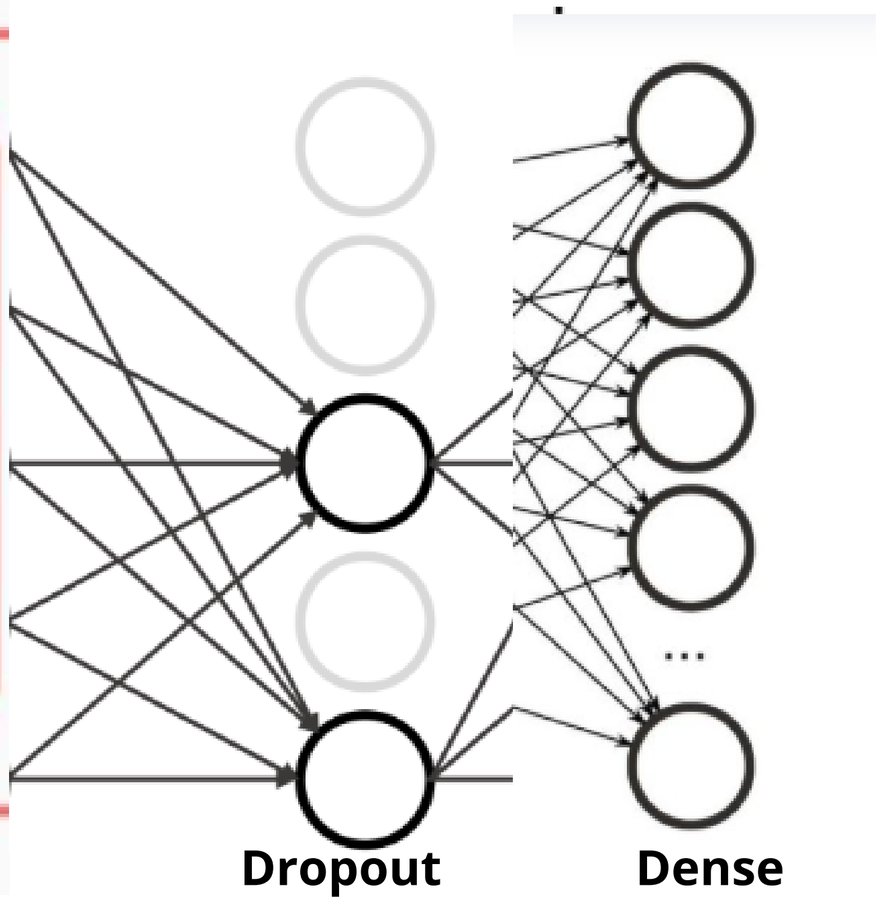


340M Parameters

Architecture



Bert



Dropout

Dense

Data

	B	C	D	E
	Questions	Category0	Category1	Category2
0	How did serfdom develop in and then leave Russia ?	DESCRIPTION	DESC	manner
1	What films featured the character Popeye Doyle ?	ENTITY	ENTY	cremat
2	How can I find a list of celebrities ' real names ?	DESCRIPTION	DESC	manner
3	What fowl grabs the spotlight after the Chinese Year of the M	ENTITY	ENTY	animal
4	What is the full form of .com ?	ABBREVIATION	ABBR	exp
5	What contemptible scoundrel stole the cork from my lunch ?	HUMAN	HUM	ind
6	What team did baseball 's St. Louis Browns become ?	HUMAN	HUM	gr
7	What is the oldest profession ?	HUMAN	HUM	title
8	What are liver enzymes ?	DESCRIPTION	DESC	def
9	Name the scar-faced bounty hunter of The Old West .	HUMAN	HUM	ind

Loss Function- Multilabel Task

MultiLabel : each sample can be in any number of the specified classes, including zero. it can be understood as a series of binary classifications:
Is sample 1 in class A – yes or no? Is sample 1 in class B – yes or no?
And so on.

Binary Cross Entropy Loss :

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

- N *is the number of samples*
- y_i i *is the true label for the* i^{th} *sample* (0 *or* 1)
- p_i *is the predicted probability that the* i^{th} *sample belongs to Positive class*

Loss Function- Multilabel Task

BCElogits Loss :

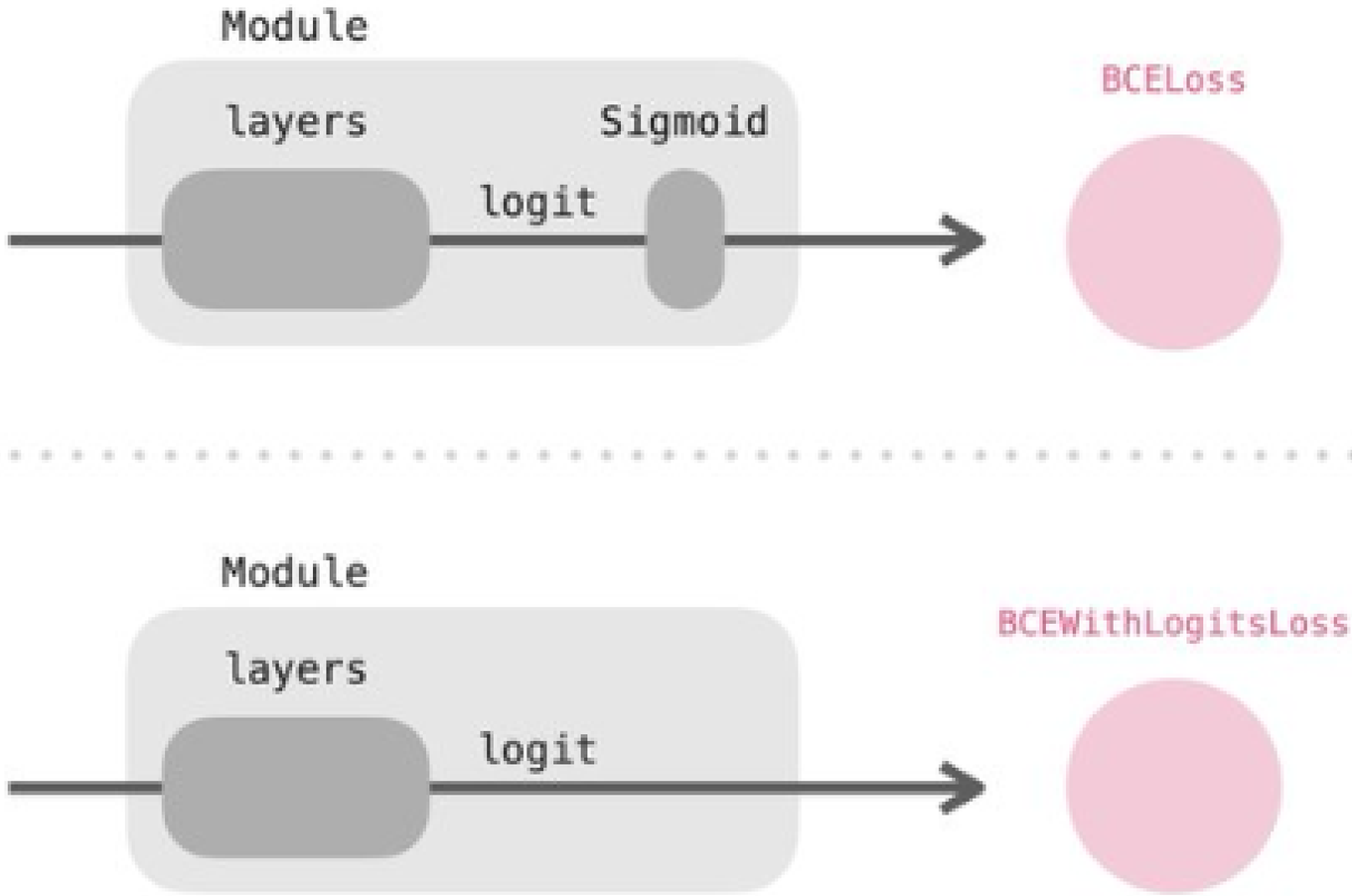
$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log_e\left(\frac{1}{1+e^{-z}}\right) + (1-y_i)\log_e\left(1-\frac{1}{1+e^{-z}}\right)\right]$$

$$\geqslant \textit{Using Quotient Rule } \log_a\left(\frac{m}{n}\right) = \log_a m - \log_a n$$

Each class label is effectively an independent binary classification problem (is this class present: yes/no?)
BCEWithLogitsLoss combines two operations efficiently:

- A sigmoid function to convert raw logits to probabilities (0-1 range)
- Binary cross entropy loss calculation

Loss Function- Multilabel Task



Conclusion

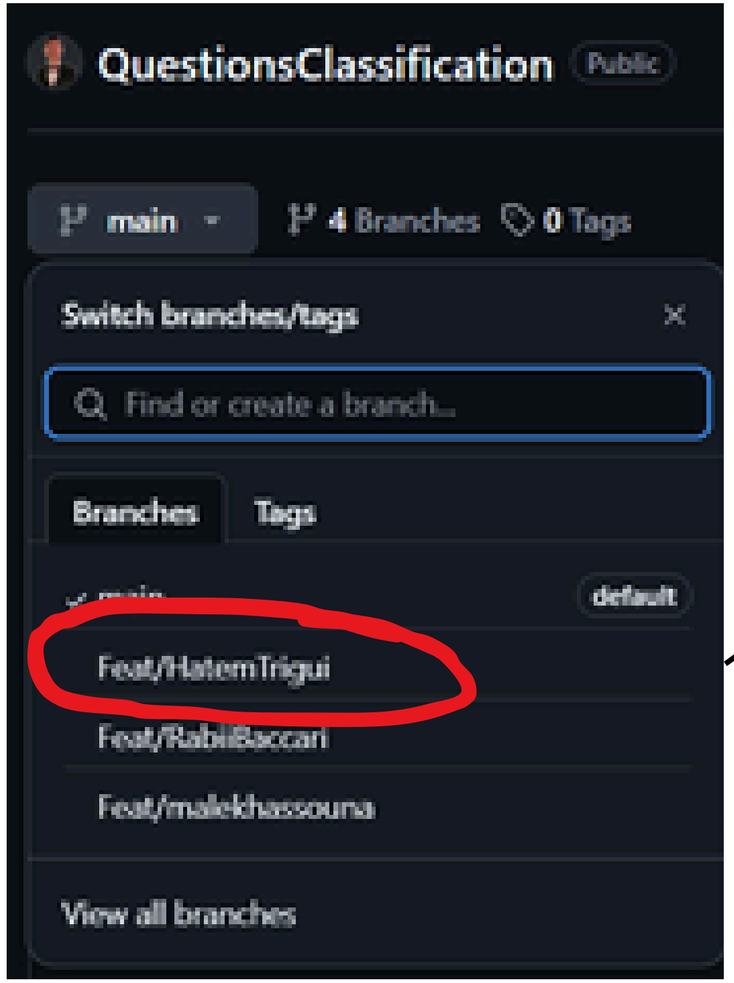
Perspectives

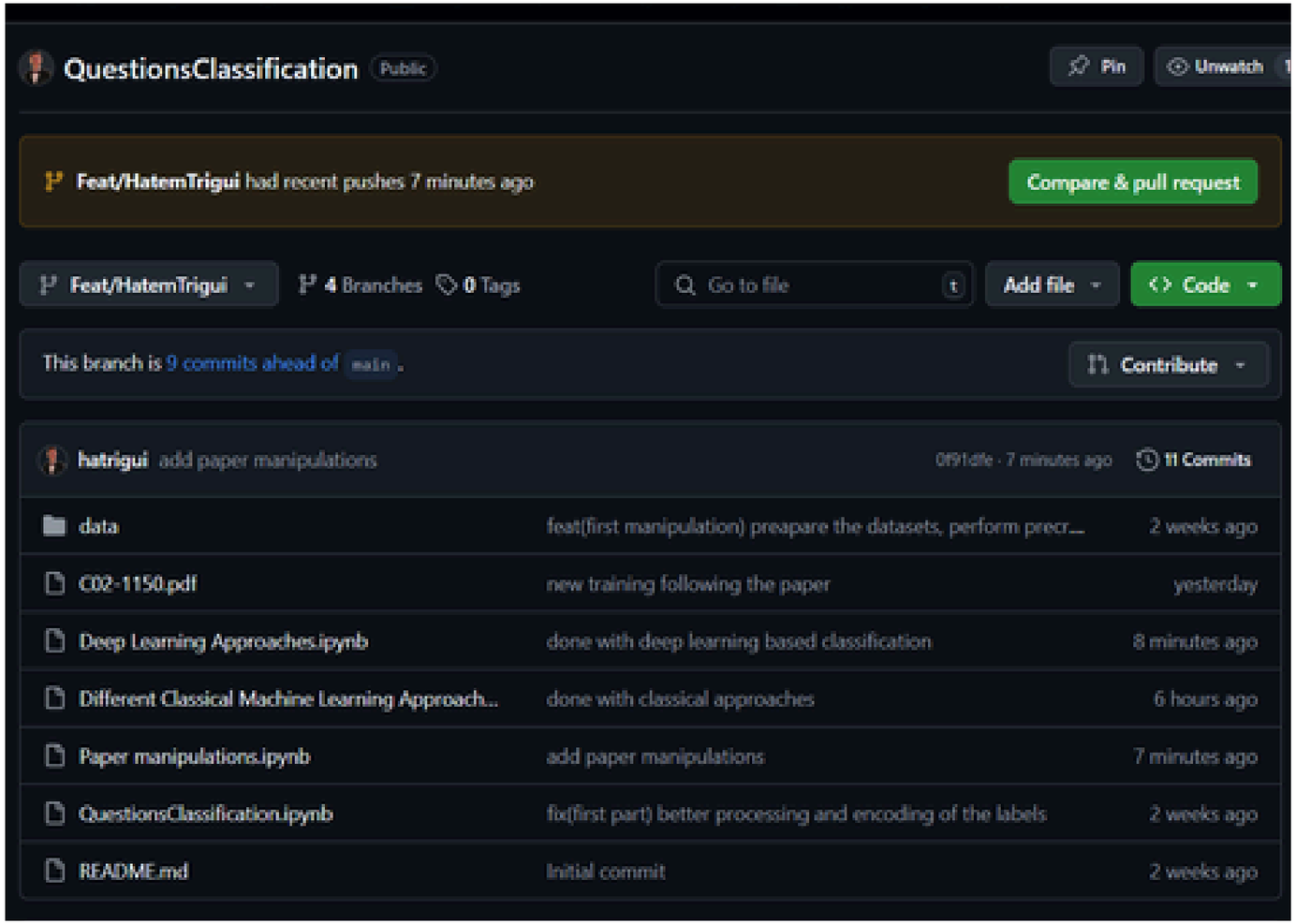
Areas of reaserch :

- Since LLMs are too costly how would Small LMs behave in this task ? like SmolLm , phi4
- LLMs being great Zero shot learners and few-shot learners ?
- How can we adress the inexplicabilty nature of Transformers/LLMs?
- Can the Small Models substitue the large ones if trained on a very specific taskn thus bringing down energy consumption and cost?

Work Methodology

version control





Github Repo Link: https://github.com/hatrigui/QuestionsClassification

THANK YOU FOR YOUR ATTENTION