Personal data detection in free text

Candidate:
Gabriele Gioetto

Supervisors:
Dr. Francesco Di Cerbo
Prof. Paolo Papotti
Prof. Giuseppe Rizzo

1859

EURECOM
Sophia Antipolis

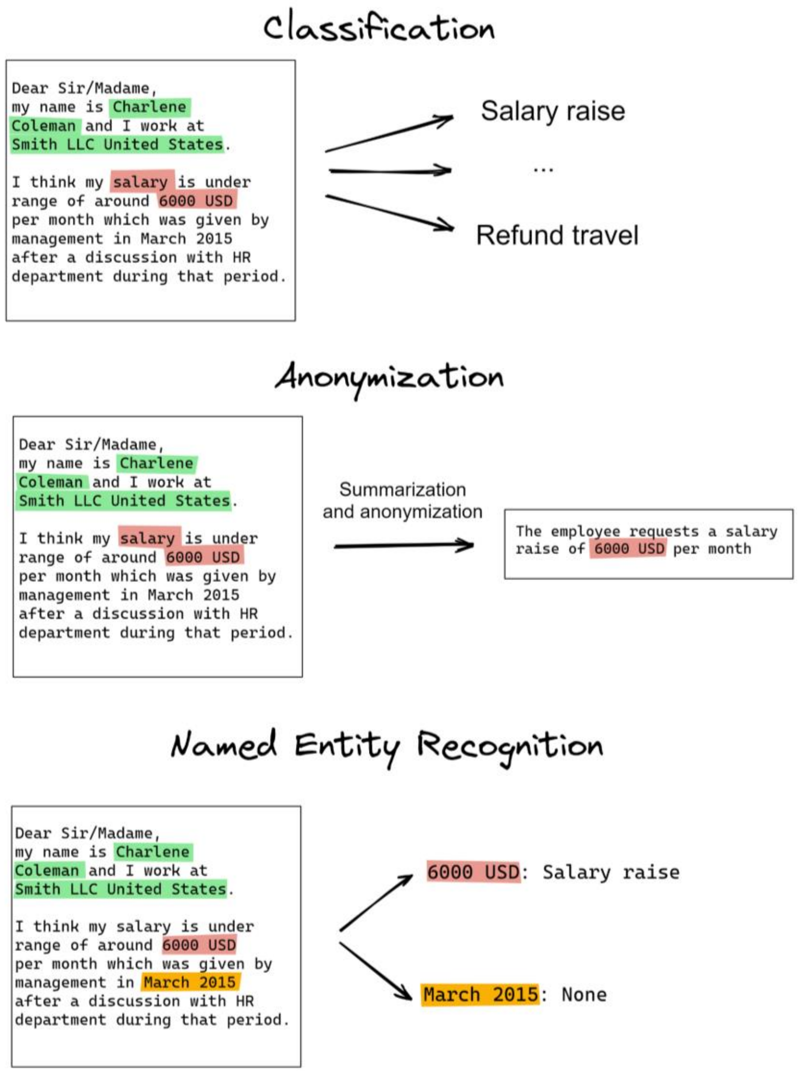
SAP

1. HR Ticket dataset generator

Dear Sir/Madame,
my name is Charlene Coleman and I work at Smith LLC United States.

I think my salary is under range of around 6000 USD per month which was given by management in March 2015 after a discussion with HR department during that period.

2. Use cases



1. HR Ticket Generation

HR Ticket Generation

- **Objective**: generate fake tickets sent to HR to train **Machine Learning** models
- Why **fake**? GDPR personal data
- **Method**: start from a dataset of **real data** to reflect real world distributions in the generated tickets

Our Approach

Creation of a Taxonomy of HR Tickets Topics

- Family situations, health conditions, work conditions (compensation)…
- Identification of key elements for each topic (e.g. type of disease, sick leave length etc)

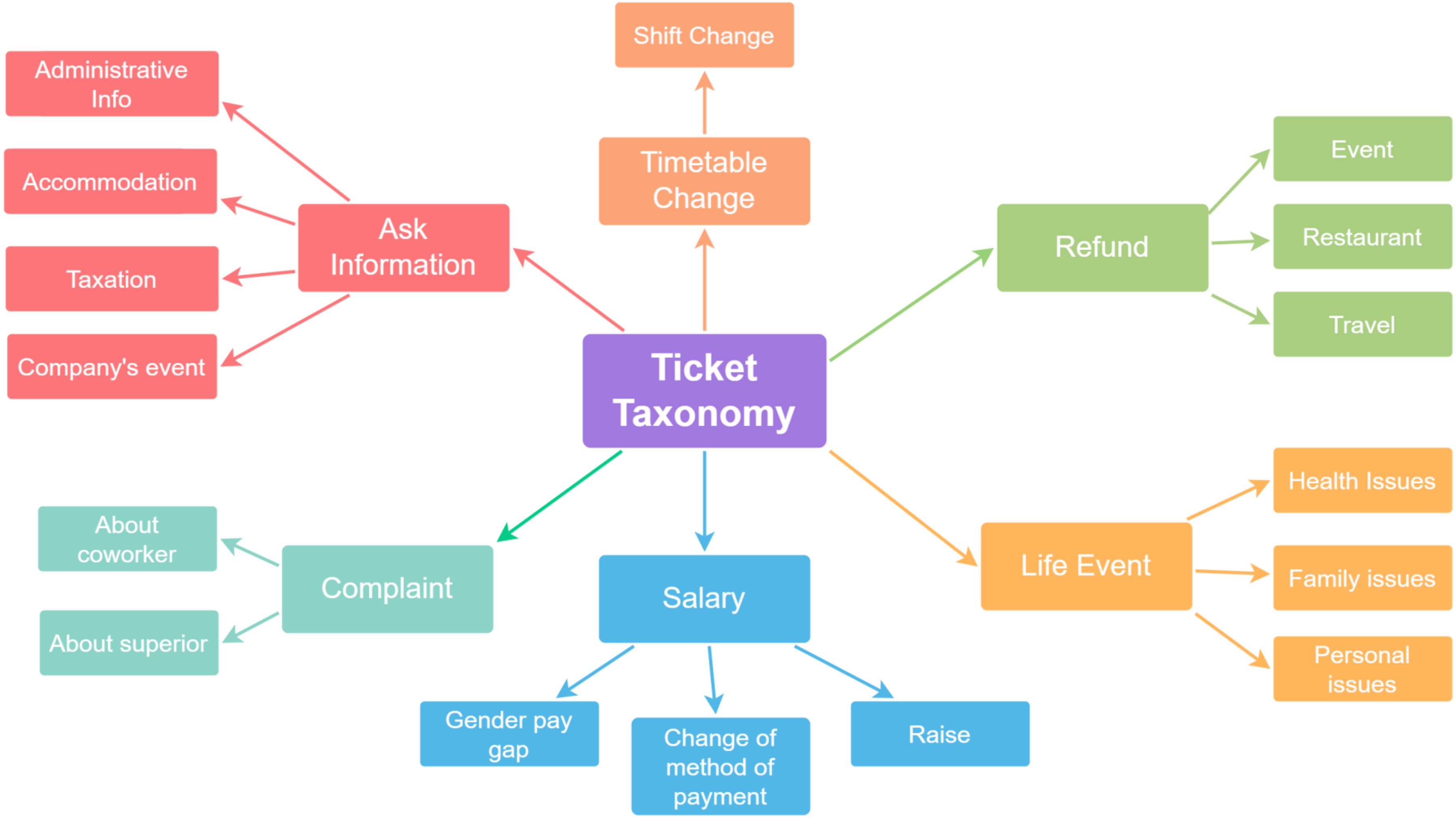
Gathering of Open Data HR information

- Public information on absenteism, discrimination at work, salary conditions ...

Ticket Generation from Open Data

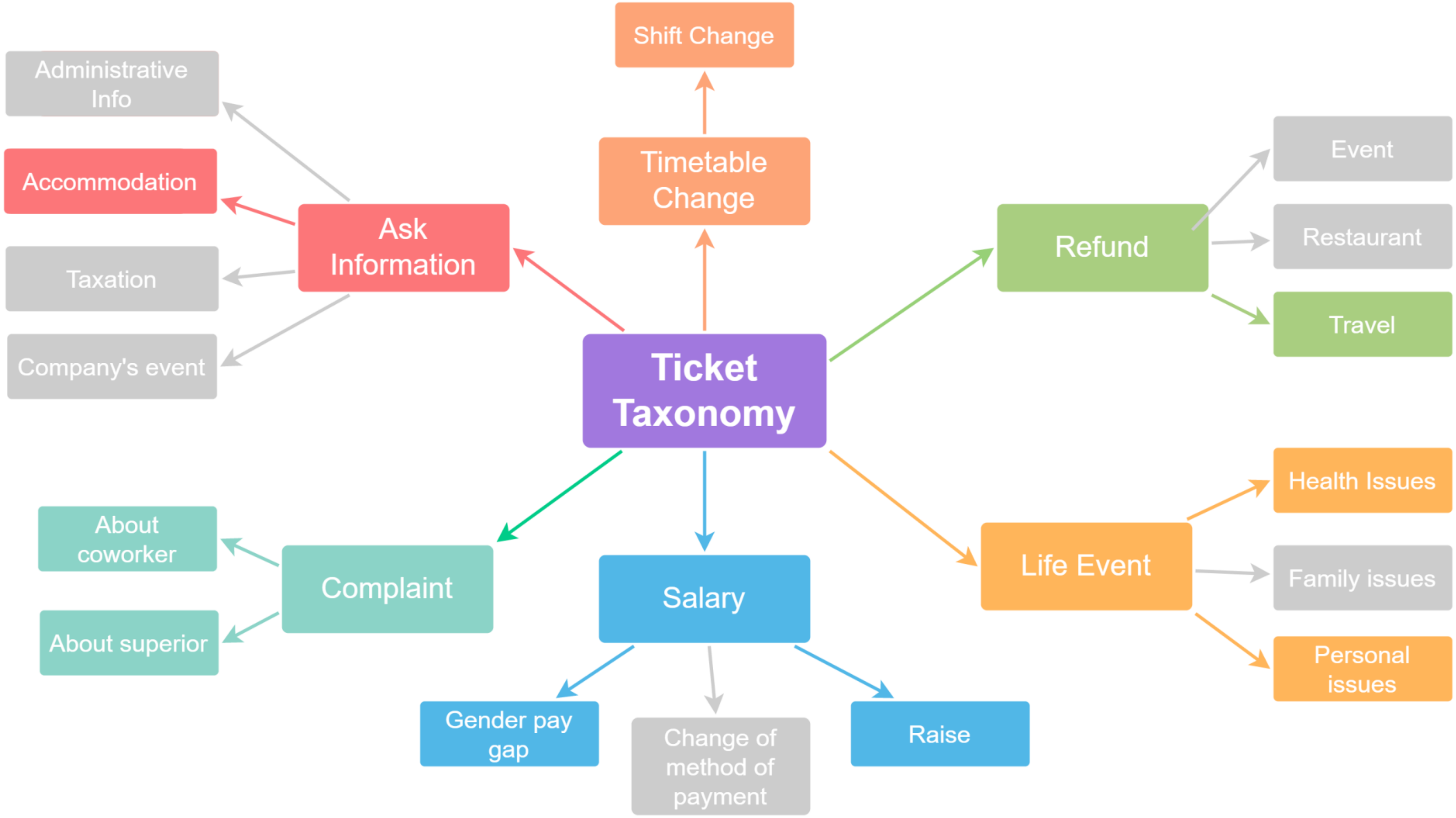
- Definition of methodology to combine key topic elements with randomly generated profile information and use all data to generate an HR ticket

Taxonomy



Ticket Taxonomy
Ask Information
Administrative Info
Accommodation
Taxation
Company's event
Timetable Change
Shift Change
Refund
Event
Restaurant
Travel
Life Event
Health Issues
Family issues
Personal issues
Salary
Gender pay gap
Change of method of payment
Raise
Complaint
About coworker
About superior

Taxonomy



Shift Change
Administrative Info
Timetable Change
Accommodation
Ask Information
Taxation
Company's event
Ticket Taxonomy
Event
Refund
Restaurant
Travel
Health Issues
About coworker
Complaint
About superior
Salary
Life Event
Family issues
Personal issues
Gender pay gap
Change of method of payment
Raise

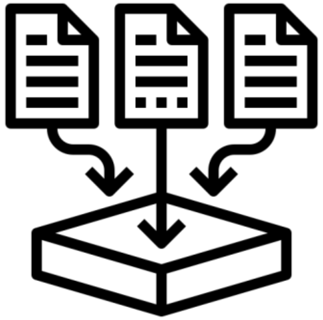
Taxonomy (Variables)

Category	Sub-category	variables
Ask Information	Accommodation	location, duration
Complaint	About coworker	complaint, reason
	About superior	complaint, reason
Timetable change	Shift change	reason_of_change, old_date, new_date
Salary	Salary raise	old_salary, new_salary, increase,work_title
	Gender pay gap	wage_gap
Life Event	Health issues	disease, number_of_days_of_sick_leave
	Personal issues	issue, number_of_days
Refund	Travel	from, to, date_travel

Method

1) Create an **employee dataset** (with random names, surnames, mails, companies, …)

name	first_name	last_name	nationality	country	email	company	company_email	ticket_date
Jivin Samra	Jivin	Samra	IN	US	jivin.samra@sathe.com	Sullivan-Byrd United States	hr@SullivanByrd.com	27/06/2018
Anna Tschentscher	Anna	Tschentscher	DE	DE	atschentscher@hauffer.com	Dobes GbR Germany	hr@DobesGbR.com	7/5/2017
Sigfried Eberth	Sigfried	Eberth	DE	US	sigfried.eberth@hoevel.de	Barrett-Davis United States	hr@BarrettDavis.com	17/01/2021
Marcia Hays	Marcia	Hays	US	DE	mhays@walsh.info	Kobelt Kostolzin AG & Co. KGaA Germany	hr@KobeltKostolzinAGCoKGaA.com	9/8/2017
Christopher Harrison	Christopher	Harrison	US	DE	charrison@sanchez-lang.net	Walter Germany	hr@Walter.com	10/10/2017



2) Create **additional employees' information** with respect to the type of ticket, following statistical distribution from real data set

name	work_title	prev_salary	standard_e	increase_	new_salary
Jivin Samra	operations specialties manager	139700	559.6	0.08	150900
Anna Tschentscher	protective service occupation	52900	320.52	0.05	55500
Sigfried Eberth	home health and personal care aides; an	30200	121.76	0.09	32900
Marcia Hays	driver/sales workers and truck driver	45300	272.76	0.06	48000
Christopher Harrison	all occupation	58200	116.52	0.06	61700

Datasets used

List of dataset used in the project:

- Absenteeism at work Data Set: records of **absenteeism at work** from July 2007 to July 2010 at a courier company in Brazil

- National Occupational Employment and Wage Estimates United States: **wage estimates** calculated with data collected from employers **in all industry sectors** in metropolitan and nonmetropolitan areas in every state and the District of Columbia

- List of events of life: list of major **events in life**

- Gender pay gap in the UK: list of UK companies with average **pay gap amongst genders**

- OpenFlights Database: datasets of **airports and flights** all over the world

- Geonames all cities with a population over 1000: datasets of all **cities** of the world with a population over 1000 people

Aggregated: the information are aggregated (Ex: average salary)

Work Title	**Annual mean wage**
Sales manager	$142,390
Cook	$29,560
…	…

Datasets

Single Record: there is a record for each employee

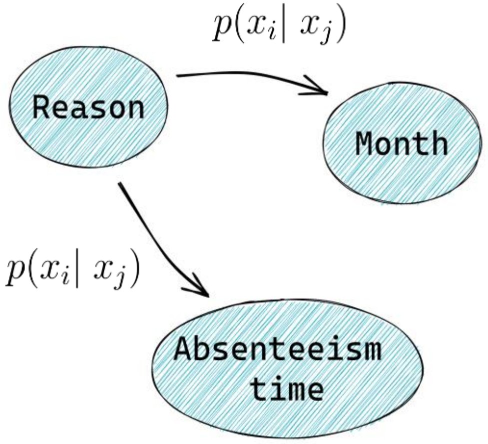
Employee ID	**Reason for Absence**	**Time off**
1	Pregnancy	120 days
2	COVID	10 days
…	…	

Privacy of datasets

Bayesian Networks

Calculate **correlations** between the attributes in the dataset, allowing us to approximate the distribution of data. Noise is then injected into each marginal to uphold the **differentially private** guarantee. We create a new dataset by sampling from these distributions.

Used when there is a **record** in the dataset **for each employee**



Inject Laplace Noise

Gaussian Noise

Add **gaussian noise** to the numerical features to augment the privacy guarantees.

Used when the data in the dataset are already **aggregated** (Ex: Mean salary)

Ticket Generation (Template)

From: ${email}
To: ${company email}
First name: ${first name}
Last name: ${last name}
Company: ${company}
Date: ${ticket date}
Ticket category: ${category}
Ticket sub-category: ${sub category}
Date start absence: ${date start absence}
Reason absence: ${reason}
Subject: Request for sick leave for ${number of days}

Dear Sir/Madame, my name is ${name} and I work at ${company}. I am requesting . I hope <generate>

Information about the employee (Generated with Faker)

Information specific to the category (Sampled from datasets)

Ticket text (Generated with generative model)

Ticket Generation (GPT-J)

GPT-J is an open source generative model developed by EletheurAI

- It achieves similar performances to GPT-3

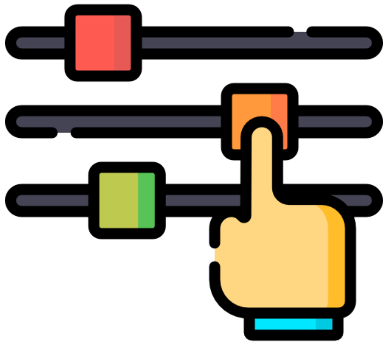
Trained on the Pile dataset

- The Pile dataset is an 825GB English text corpus composed by Academic, Internet, Prose, Dialogue and Misc categories

Two main differences with GPT-3

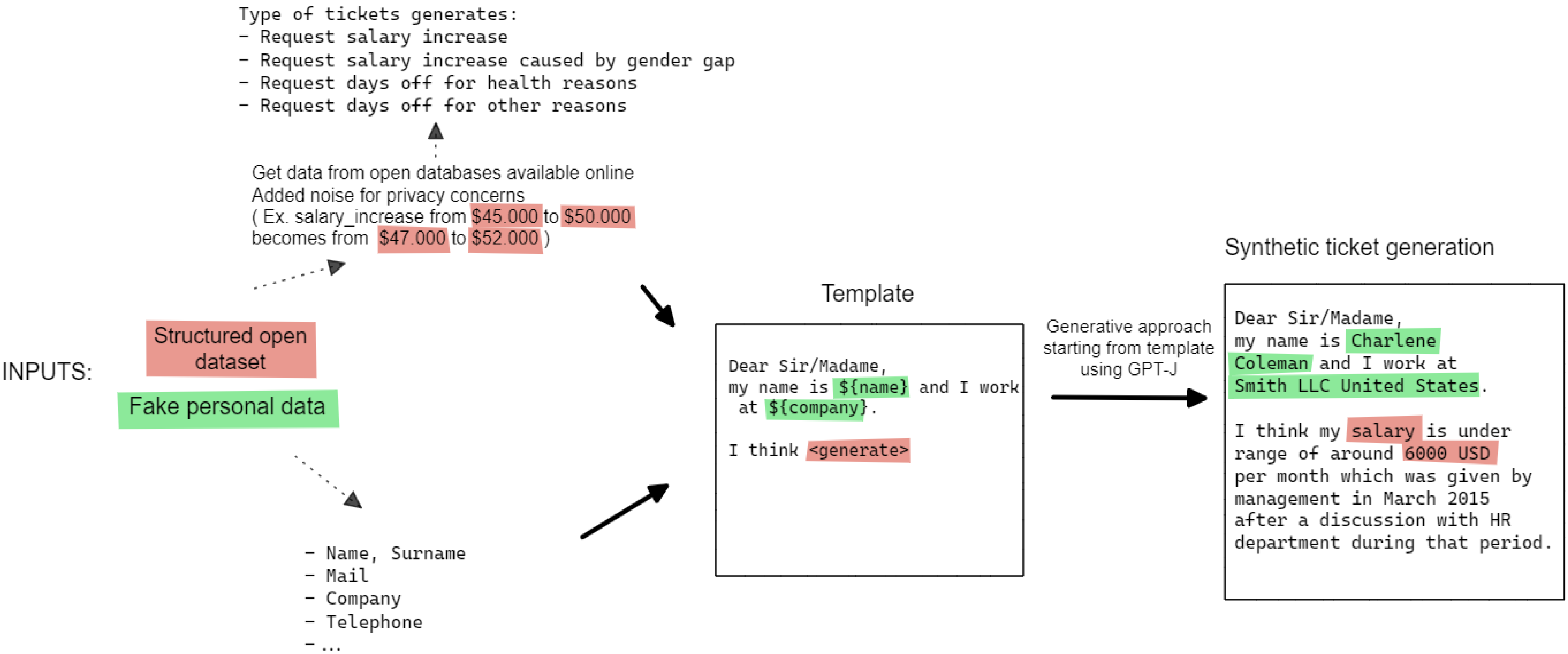
- Rotary Embedding
- Attention layer and feed forward layer in parallel

Ticket Generation (Parameters)



- Min length
- Max length
- Top k
- Top p
- Temperature
- Repetition penalty
- Length penalty
- No repeat ngram size
- Num beams
- Do sample
- Bad words
- Force words

Ticket Generation (Complete Schema)



Ticket Generation (Survey)

Why?

- To understand if the tickets generated by the model were **similar** to real tickets

How?

- We gave each respondent a excel file with **20 random prompts** to fill

No personal data

- Each prompt had some info attached. We asked the respondent to use the info to write the ticket in a way that felt natural to them

Ticket Generation (Survey)

Ticket details		**Detail used (X if yes, empty if not)**
First name	Zachary	
Last name	Traore	
Ticket category	Life event	
Ticket sub-category	Health issues	
Date start absence	18/07/2017	
Reason absence	COVID	X
Subject	Request for sick leave for 2 days	

Ticket text

Dear Sir/Madame, I am positive to COVID and I cannot come to work for the next days. I don't exactly know when I will be able to return to work. Zachary

Ticket Generation (Results)

- **TTR**

$$TTR = \frac{number\_of\_unique\_bigrams}{total\_number\_of\_bigrams}$$

- **NOUN RATIO**

$$noun\_ratio = \frac{counter\_noun\_words}{counter\_all\_words}$$

- **VERB RATIO**

$$verb\_ratio = \frac{counter\_verb\_words}{counter\_all\_words}$$

- **WORD FREQUENCY**

$$word\_freq = \begin{cases} log_e(word\_freq\_wiki), & \text{if } word\_freq\_wiki \geq 10 \\ skip, & \text{otherwise} \end{cases}$$

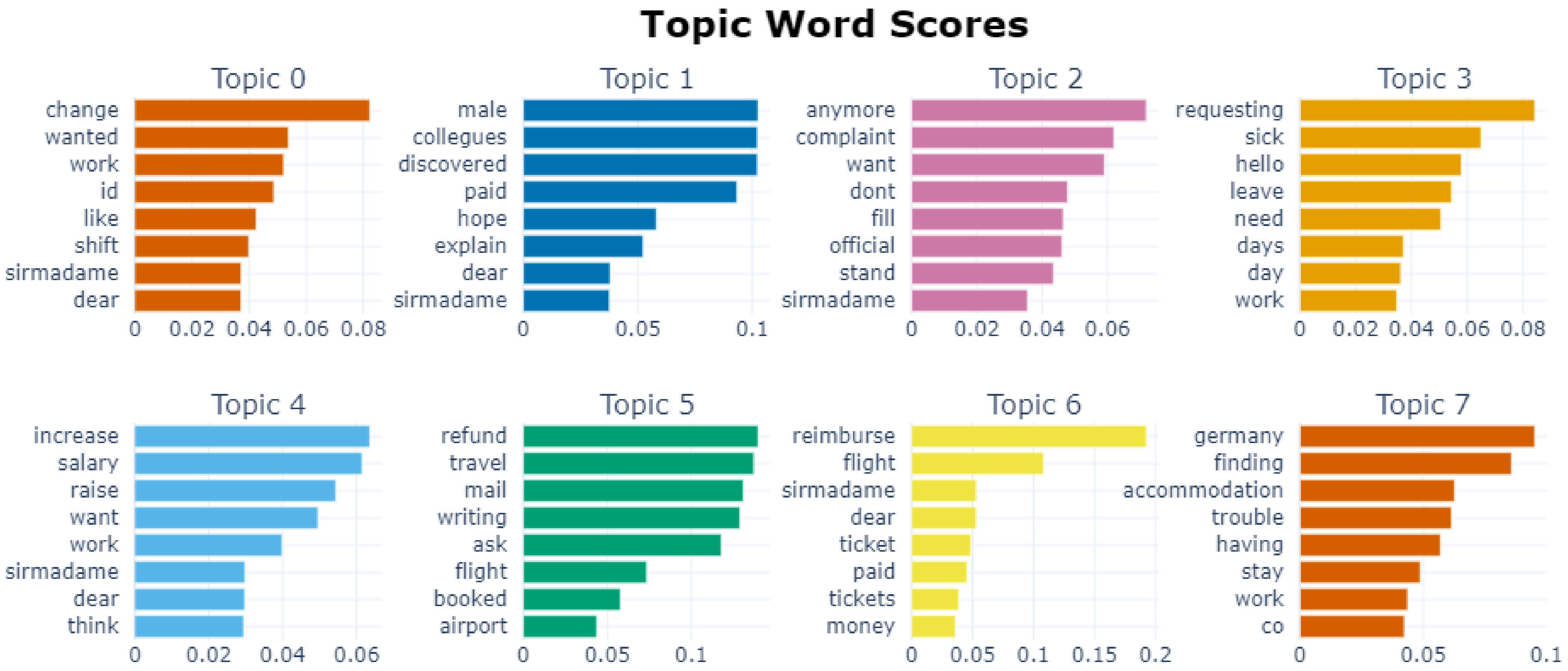
Ticket Generation (Results)

Main Takeaways:

- The model tends to write **more unique words** and not to repeat itself
- The model tends to use **more nouns** than a real person
- The model tends the **same number of verbs** as a real person
- The average **number of words** (word count) **varies a lot from category to category**, both in generated and survey tickets

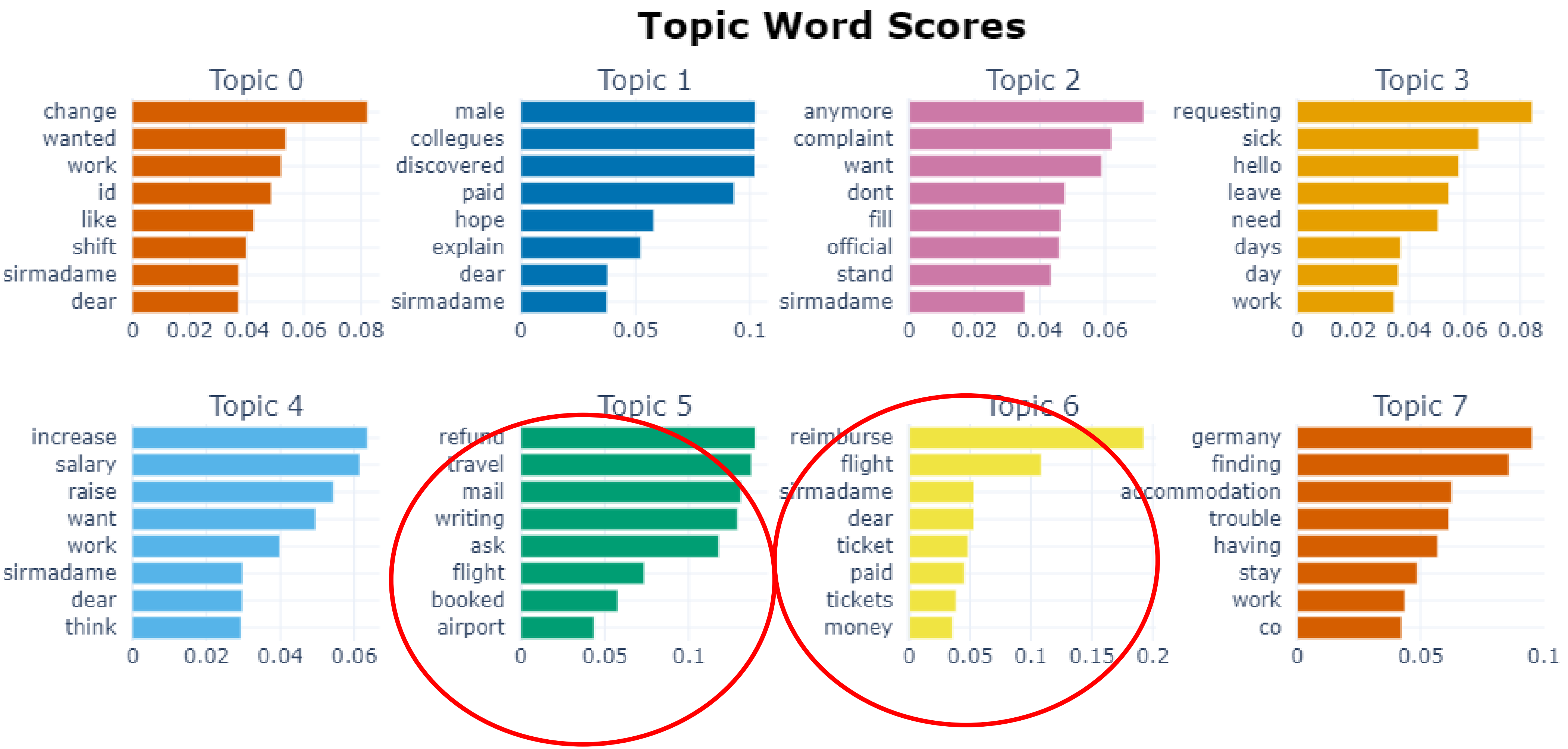
Ticket Generation (Results)

BERTopic



Ticket Generation (Results)

BERTopic



Missing category: Request of time off due to personal reasons

2. Use Cases

USE CASES

WHY?

- In order to assess the **suitability** of the HR ticket dataset for downstream ML tasks

HOW?

- **Training set**: HR ticket dataset generated by our application
- **Test set**: HR tickets gathered with survey

TASKS

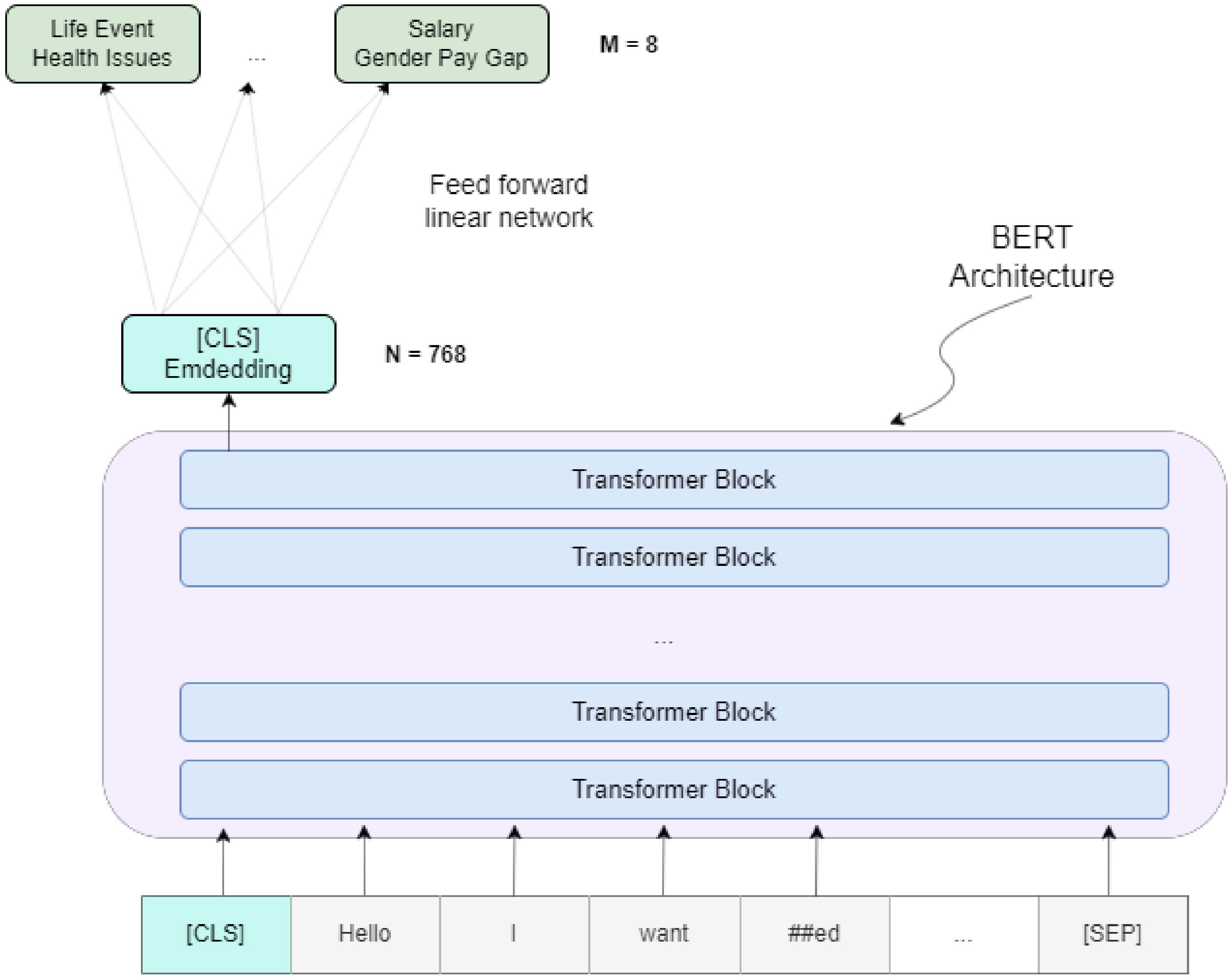
- Classification
- Anonymization
- Named Entity Recognition

Classification

Classification of the tickets' categories

Two models:

- *BERT*
- *FastText*



Classification (Results)

F1 score: ***0.78***

Hyperparameters:

- *Epochs: 5*
- *Learning rate: 5e-05*
- *Weight decay: 0.001*
- *Batch size: 8*
- *Warmup steps: 500*

Anonymization

Classical anonymization

The new intern at my office, the one with red hair, caught covid last week

↓ *Presidio*

The new intern at my office, the one with red hair, caught covid <DATE_TIME>

Our anonymization

The new intern at my office, the one with red hair, caught covid last week

↓

An employee was sick last week

Anonymization

Dear Sir/Madame,
my name is Charlene Coleman and I work at Smith LLC United States.

I think my salary is under range of around 6000 USD per month which was given by management in March 2015 after a discussion with HR department during that period.

Summarization and anonymization

Using a few-shot approach with model T5, removing personal info without affecting the content needed for analysis matters

The employee requests a salary raise of 6000 USD per month

Evaluation of the summarization and anonymization of the tickets using:
- question-answering models
- cosine-similarity

Anonymization (Evaluation: QAGS)

- QAGS

Dear Sir/Madame,
my name is Charlene Coleman and I work at Smith LLC United States.

I think my salary is under range of around 6000 USD per month which was given by management in March 2015 after a discussion with HR department during that period.

Summarization

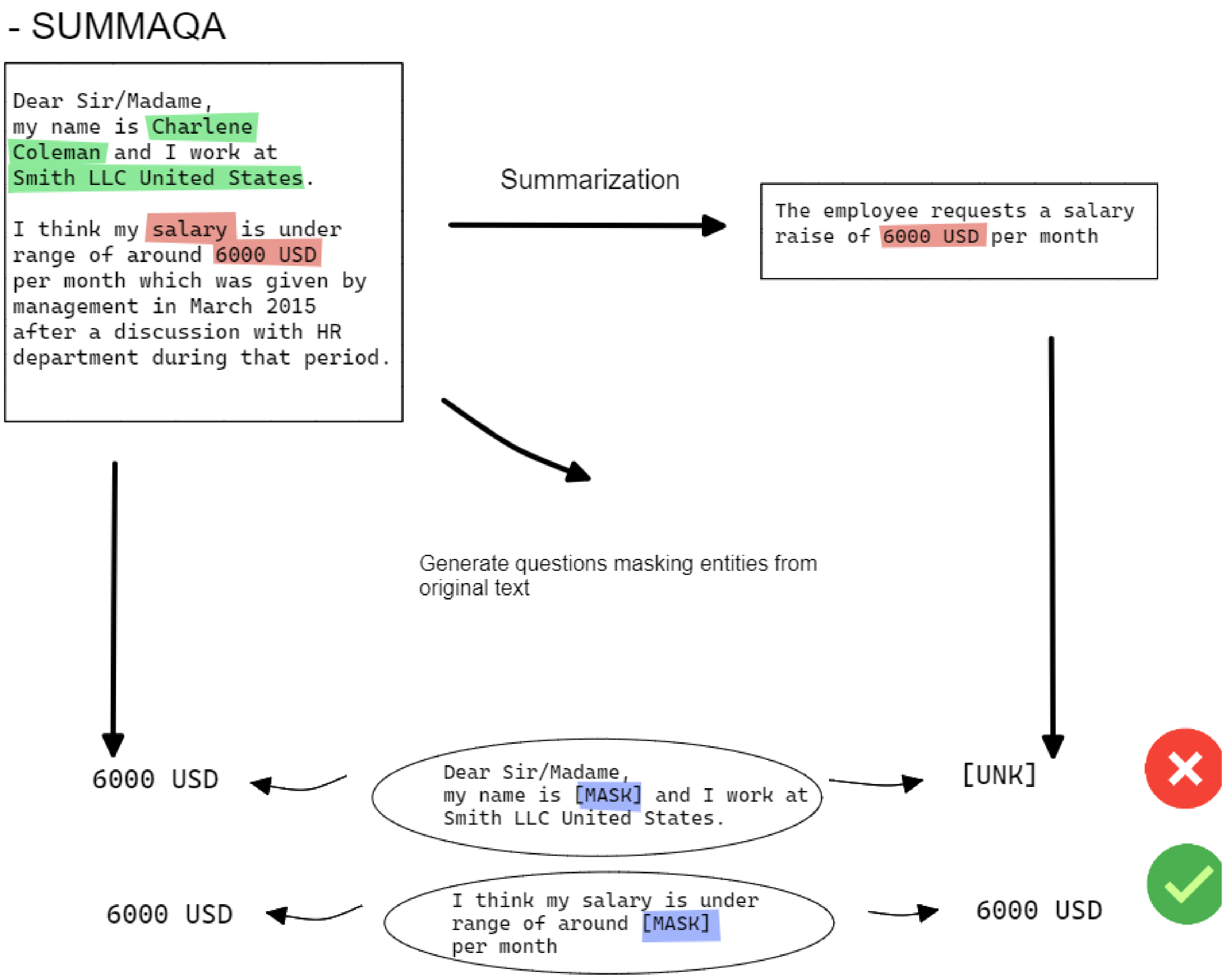
The employee requests a salary raise of 6000 USD per month

Generate questions using as answers entities found from summary (ex. 6000 USD, employee...) and as context the summary. Questions generated in an Unsupervised manner, using a finetuned BART

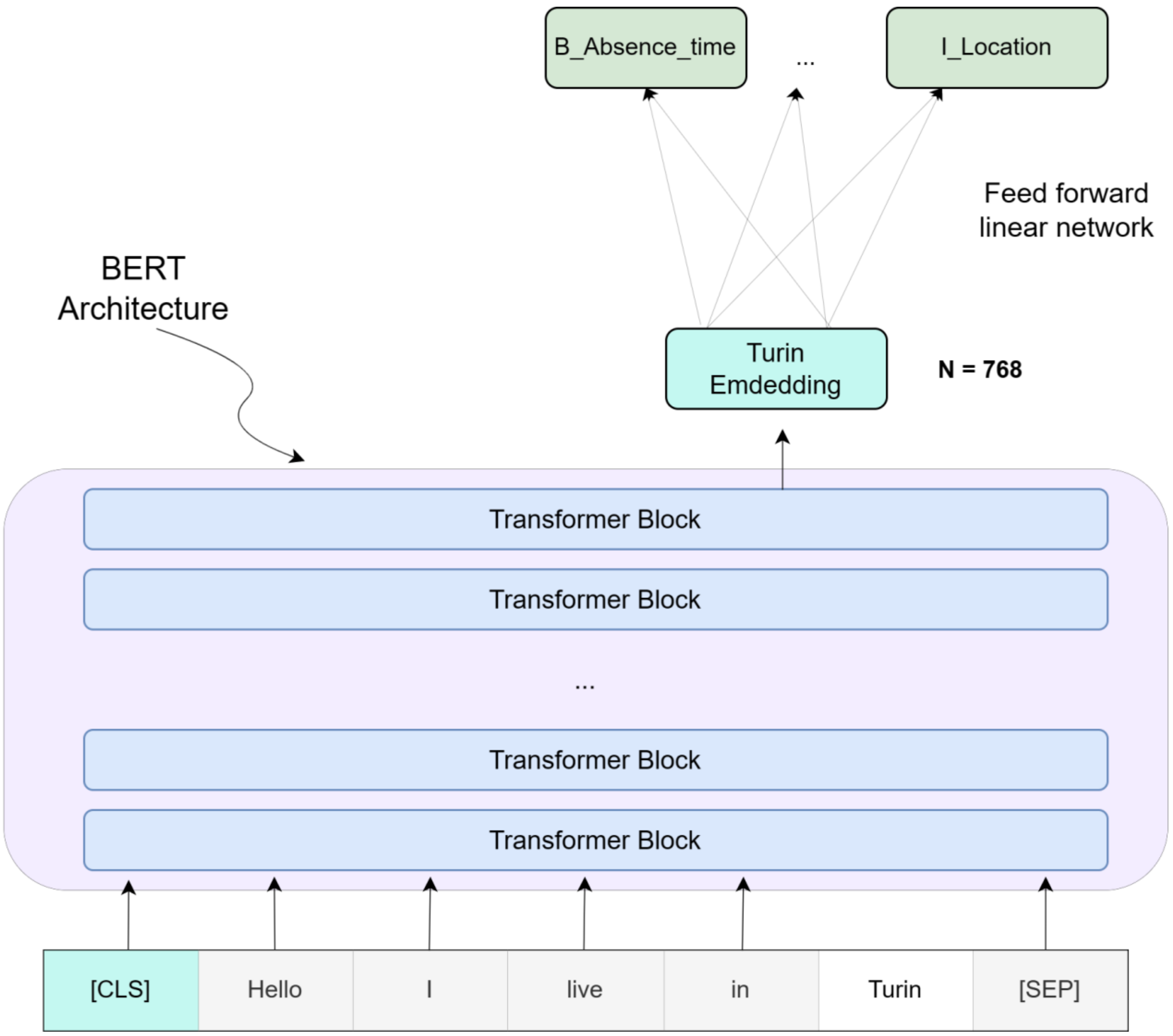
6000 USD ← What salary raise the employee asked? → 6000 USD

Charlene Coleman ← What's the employee's name? → [UNK]

Anonymization (Evaluation: SummaQA)



Named Entity Recognition (Classical)



NER on the tickets' entities (the entities are the variables taken from the datasets)

Main Problem:
The model could not recognize from entites with **almost identical format** but used in **different context** (Ex: a date of start absence , entity belonging to the requests of time off due to health reasons, and a date of travel, entity belonging to requests of refund for travel).

Named Entity Recognition (with Spacy pre-trained model)

Look for entities with Spacy

Dear Sir/Madame,
my name is Patrick
Lagarde. Could you help
me finding a nice place
to stay for the next
twelve month in Tours
(France)? The cost will
be around 300$ per
month depending on the
house type.

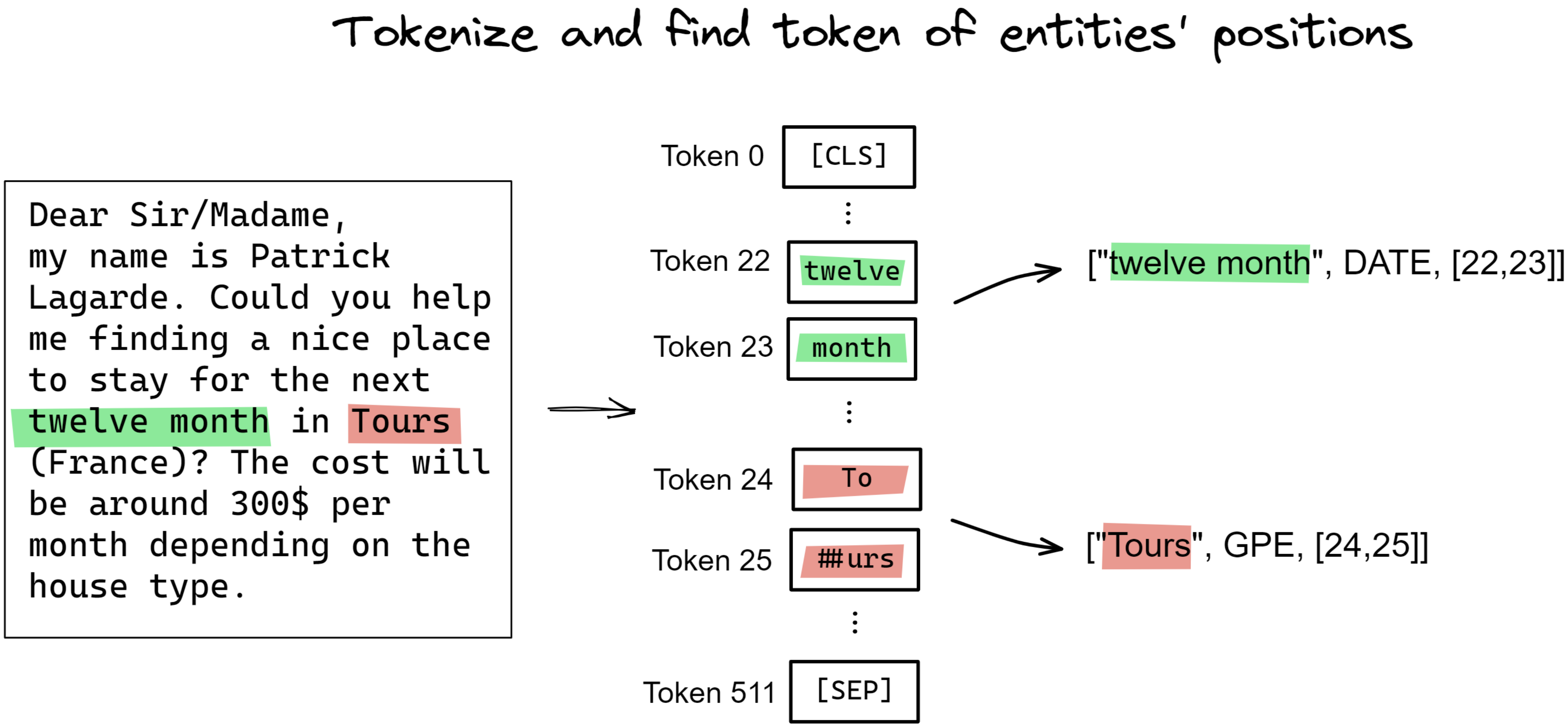
["twelve month", DATE, 52, 64]

["Tours", GPE, 52, 64]

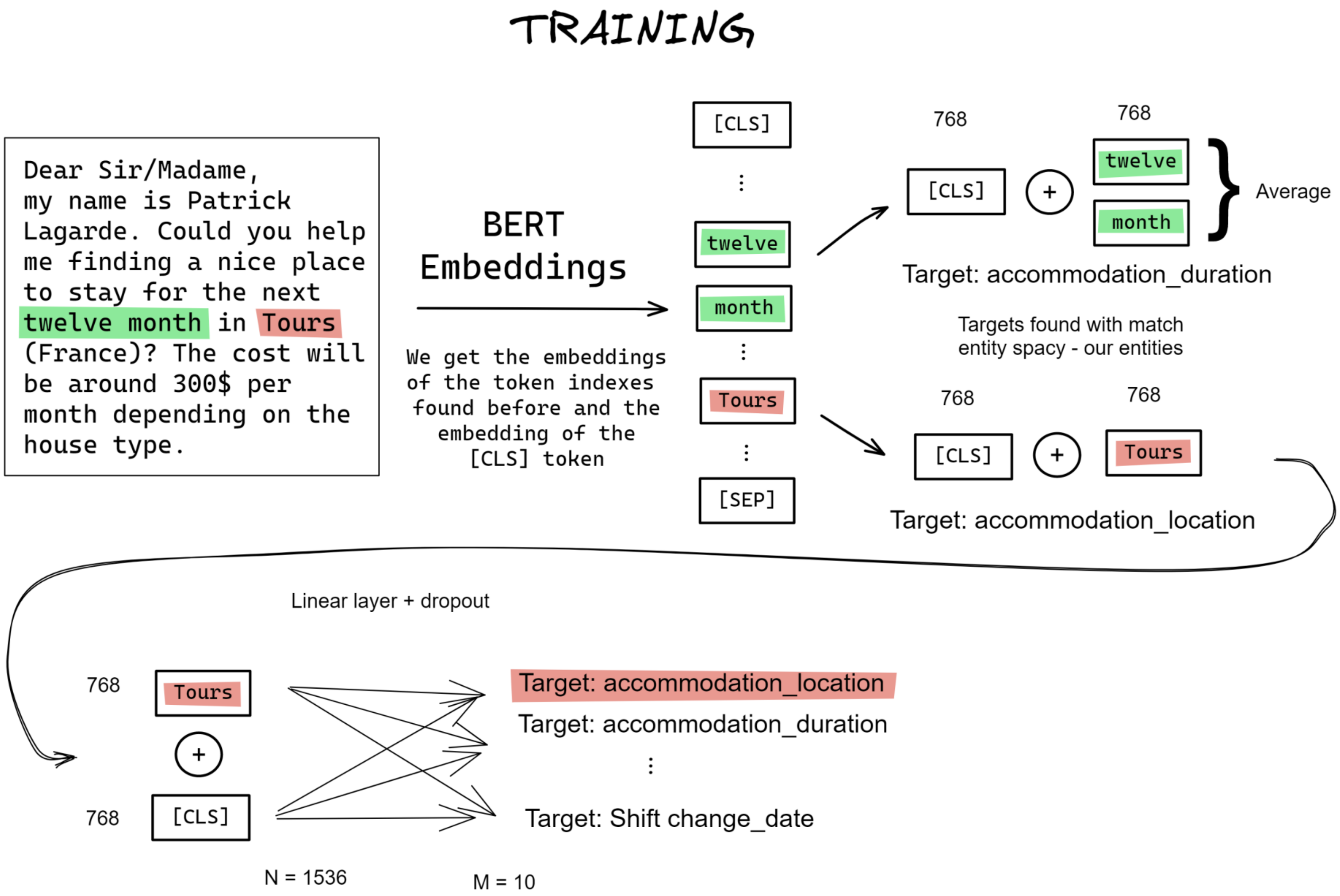
~~["300$", MONEY, 106,110]~~

filtered out

Named Entity Recognition

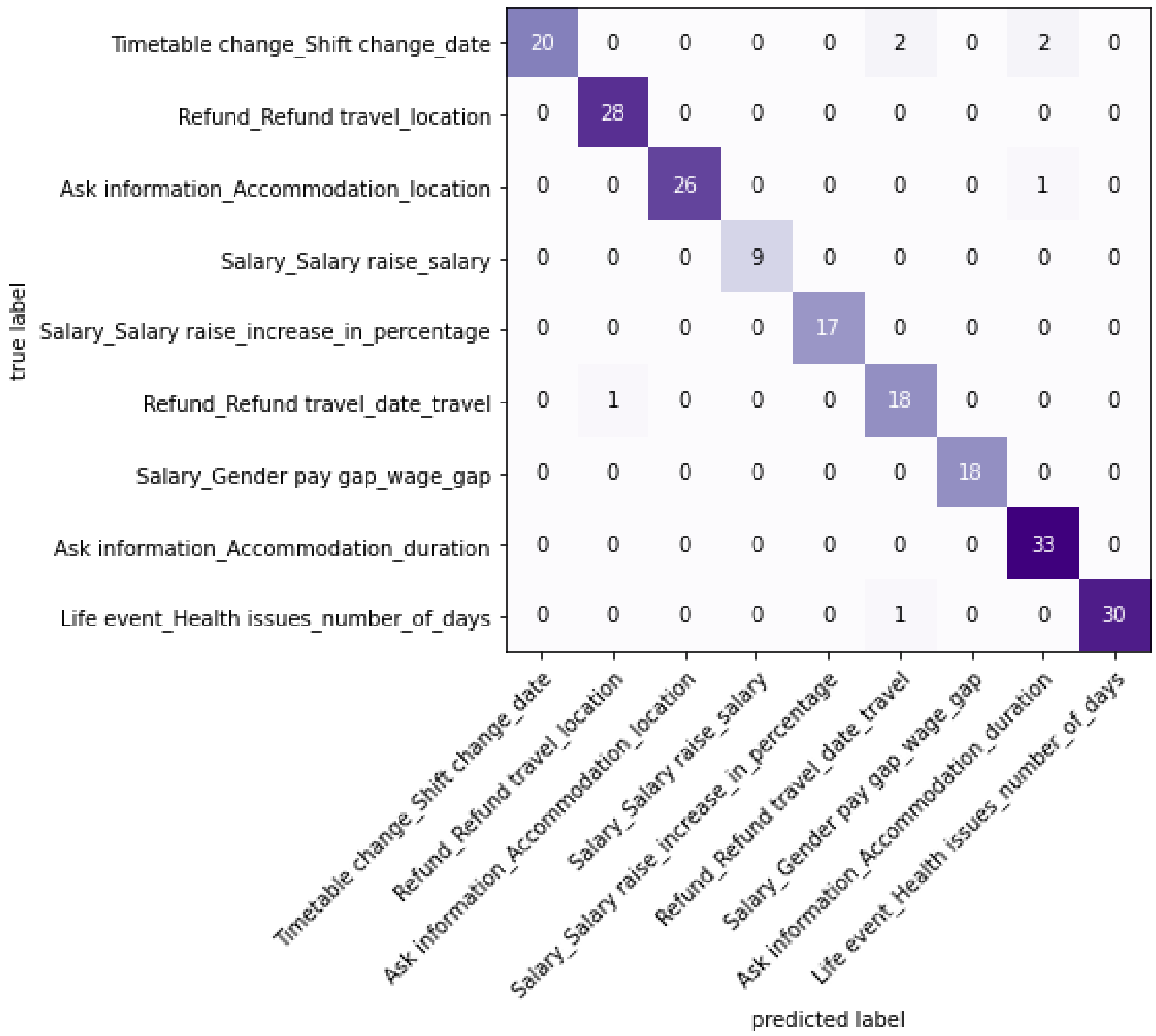


Named Entity Recognition



Results

F1 score: ***0.96***



Thanks for the attention!