Natural Language Processing for Trend Forecasting

Heládio Sampaio Lopes

Computer Engineering (ITA 2020)

CONTENT

CONTENT

INTRODUCTION

Over the years, more and more knowledge is generated and we humans are not able to process such an amount of information. Natural language processing emerges as a technology capable of assisting us in this hard task.

INTRODUCTION

Over the years, more and more knowledge is generated and we humans are not able to process such an amount of information. Natural language processing emerges as a technology capable of assisting us in this hard task.

CONTENT

LITERATURE TO REVIEW | Text Processing Techniques

Natural Language Processing is an artificial intelligence branch capable of making computers understand and extract information from human language. Identify different topics in a document's set, classify texts on predefined topics and extract sentiment are some examples of task that could be done with NLP.

Normalization

Tokenization

Bag of Words

TF-IDF

LITERATURE TO REVIEW | Text Processing Techniques

Natural Language Processing is an artificial intelligence branch capable of making computers understand and extract information from human language. Identify different topics in a document's set, classify texts on predefined topics and extract sentiment are some examples of task that could be done with NLP.

Normalization

Tokenization

Bag of Words

TF-IDF

LITERATURE TO REVIEW | Text Processing Techniques

Natural Language Processing is an artificial intelligence branch capable of making computers understand and extract information from human language. Identify different topics in a document's set, classify texts on predefined topics and extract sentiment are some examples of task that could be done with NLP.

Normalization

Tokenization

Bag of Words

TF-IDF

0: Something in the way she moves
1: Something in her smile she knows
2: Something in the things she shows me

Something, in, the, way, she, moves, her, smile, knows, things, shows, me

LITERATURE TO REVIEW | Text Processing Techniques

Natural Language Processing is an artificial intelligence branch capable of making computers understand and extract information from human language. Identify different topics in a document's set, classify texts on predefined topics and extract sentiment are some examples of task that could be done with NLP.

Normalization

Tokenization

Bag of Words

TF-IDF

{ Something, in, the, way, she, moves, her, smile, knows, things, shows, me }

	Something	In	The	Way	She	Moves	Her	Smile	Knows	Things	Shows	Me
0	1	1	1	1	1	1	0	0	0	0	0	0
1	1	1	0	0	1	0	1	1	1	0	0	0
2	1	1	1	0	1	0	0	0	0	1	1	1

LITERATURE TO REVIEW | Text Processing Techniques

Natural Language Processing is an artificial intelligence branch capable of making computers understand and extract information from human language. Identify different topics in a document's set, classify texts on predefined topics and extract sentiment are some examples of task that could be done with NLP.

Normalization

Tokenization

Bag of Words

TF-IDF

$$w_{i,j} = \mathrm{tf}_{i,j} \times \log\left(\frac{N}{\mathrm{df}_i}\right)$$

	T_1	T_2	T_3	...	T_N
(1)					
(2)					
(3)					

LITERATURE TO REVIEW | Word Embedding

Natural Language Processing is an artificial intelligence branch capable of making computers understand and extract information from human language. Identify different topics in a document's set, classify texts on predefined topics and extract sentiment are some examples of task that could be done with NLP.

LITERATURE TO REVIEW | Topic Clustering

Natural Language Processing is an artificial intelligence branch capable of making computers understand and extract information from human language. Identify different topics in a document's set, classify texts on predefined topics and extract sentiment are some examples of task that could be done with NLP.

CONTENT

RELATED WORKS

Finding meaningful topics in a document collection has been used for a lot of authors for the most various applications like forecast trends. Predicting future trends can be very helpful in various applications, like to model the evolution of research.

Topic Modeling

Jelodar et al (2020) recently use topic modeling in Reddit related posts about the new disease Covid-19 to group similar comments and perform a sentiment analysis.

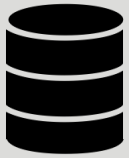
Trend Forecast

Using pre labeled scholarly articles over 25 years, Shen (2018) performed a neural network forecaster to study the topics growth and codependency between them.

Hurtado et al (2016) use topic modeling to inspect research publications to model the evolution of the direction of research and forecast the near future trends in IT industry.

RELATED WORKS

Hurtado *et al* (2016) use topic modeling to inspect research publications, patents, and technical reports aiming to model the evolution of the direction of research and forecast the near future trends in IT industry.



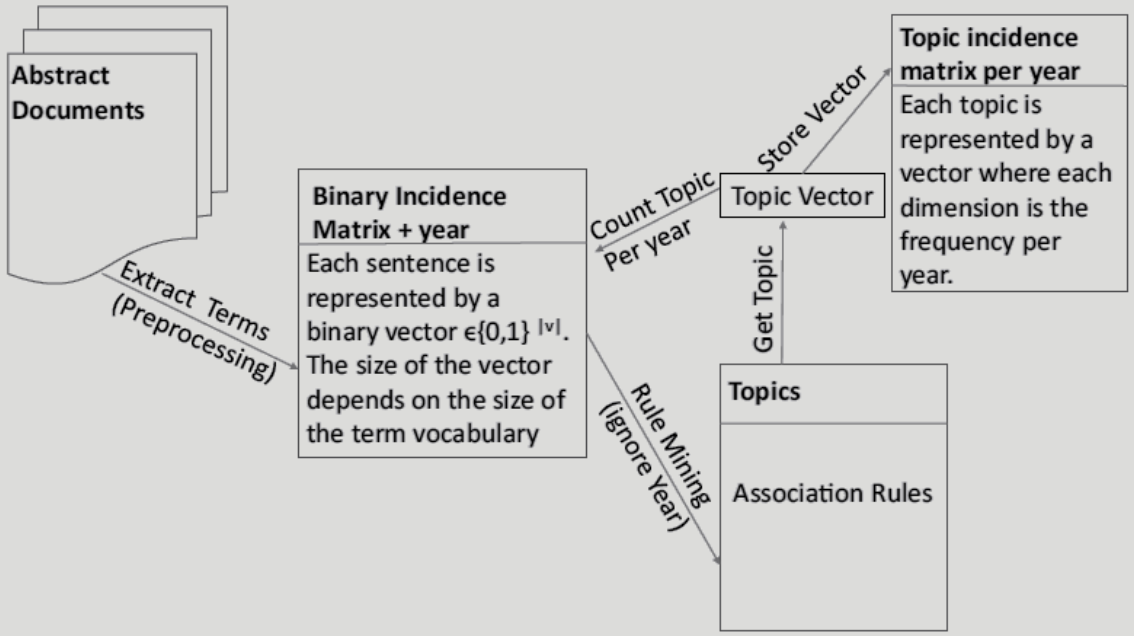
Data set containing titles and abstracts over more than six thousand academic papers distributed between 2002 and 2010, data acquired by Tang *et al.* (2008).

Sentence-level association aiming to discover meaningful topics to study the temporal correlation between the topic and predict the popularity of research topics in the future.

Find topics in the set of articles, model them as an annual time series, forecast those series using the others as input and evaluate the prediction.

RELATED WORKS

Hurtado *et al* (2016) use topic modeling to inspect research publications, patents, and technical reports aiming to model the evolution of the direction of research and forecast the near future trends in IT industry.



- Split the documents in transaction level
- Preprocess the transactions:
  - Basic Normalization
  - Remove common words in scientific publications
- Slight variation of BoW
- Topic discovery and rule refinement
- Topic incidence matrix per year

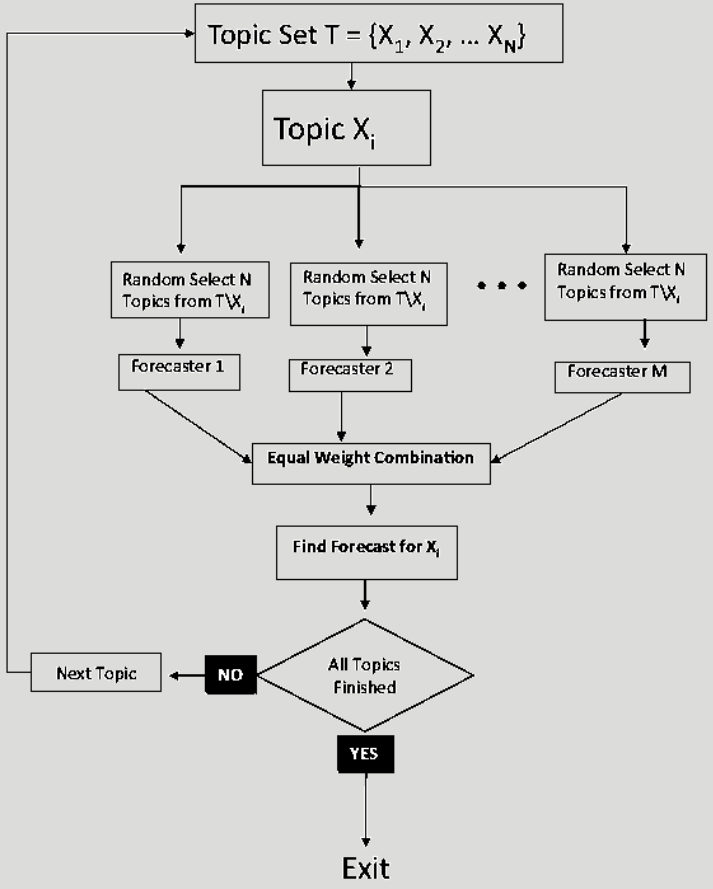
RELATED WORKS

Hurtado *et al* (2016) use topic modeling to inspect research publications, patents, and technical reports aiming to model the evolution of the direction of research and forecast the near future trends in IT industry.

Topic	2002	2003	2004	2005	2006	2007	2008	2009	2010
Random_walk	0	0	4	4	4	10	9	11	19
Neural_network	14	4	2	2	2	5	7	12	7
Compon_princip	2	3	9	5	12	8	11	7	9
Transfer_learn	0	0	0	0	1	6	10	19	18
Collabor_filter	0	3	10	9	2	8	19	16	20
Select_featur	18	9	28	35	17	32	47	35	77
Topic_model	4	0	5	0	22	27	22	56	36

RELATED WORKS

Hurtado *et al* (2016) use topic modeling to inspect research publications, patents, and technical reports aiming to model the evolution of the direction of research and forecast the near future trends in IT industry.

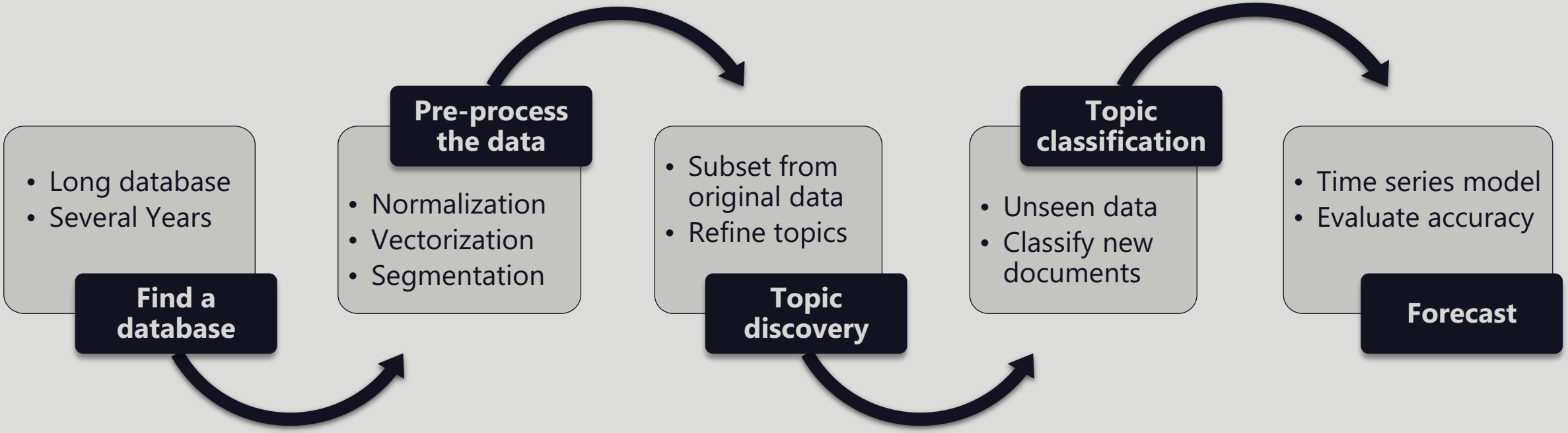


- Iterate over the topic set
- Ensemble topic forecast
- Predict a topic with others chosen at random
- Average the ensemble
- Evaluate:
  - MSE
  - Accuracy

CONTENT

MATERIALS AND METHODS | Objectives

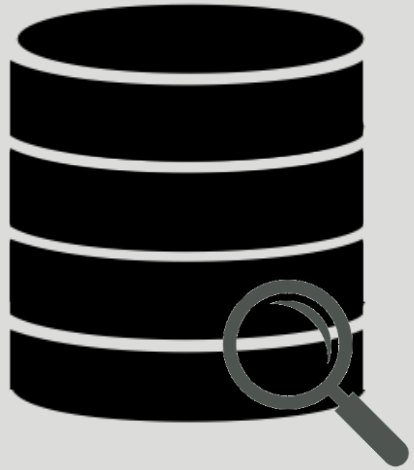
As discussed earlier, we want to build models capable of make predictions regarding the evolution of discovered topics in a set of documents and identify the discovered topics in real time.



MATERIALS AND METHODS | Database

As discussed earlier, we want to build models capable of make predictions regarding the evolution of discovered topics in a set of documents and identify the discovered topics in real time.

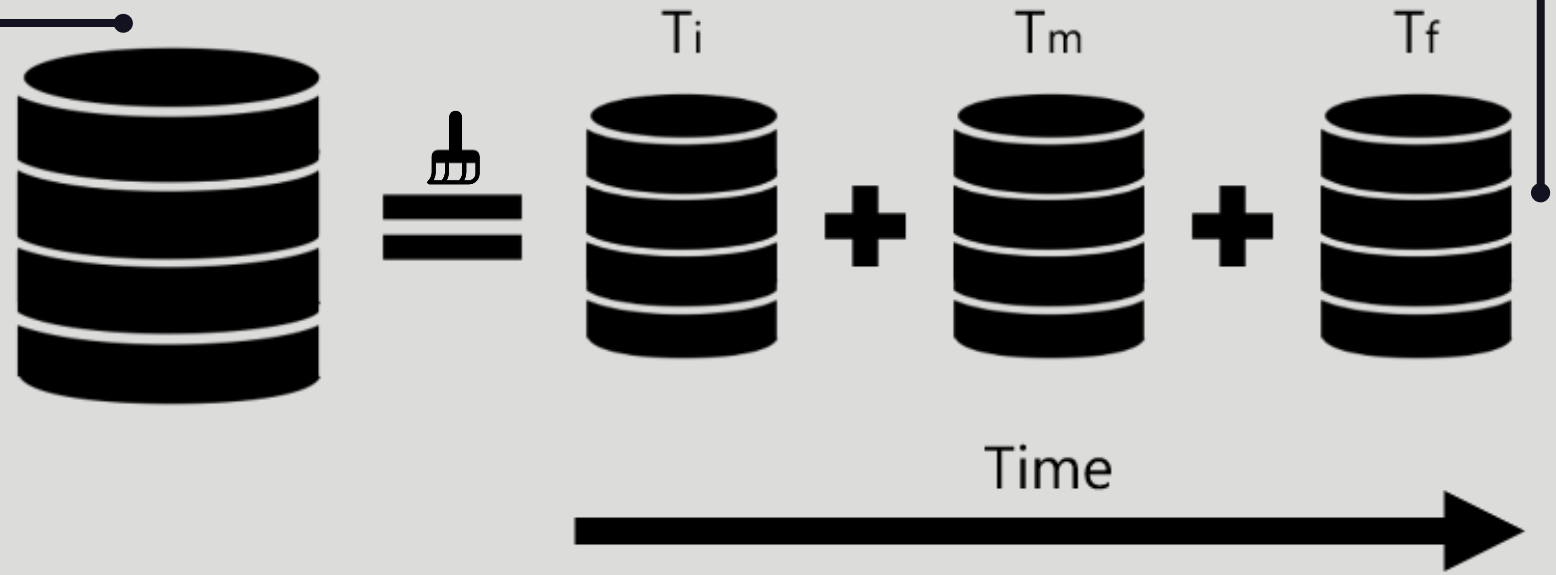
- Wikipedia Daily News
- Newspapers Articles
- Academic Papers
- Social Media - Reddit



MATERIALS AND METHODS | Pre-processing the Data

As discussed earlier, we want to build models capable of make predictions regarding the evolution of discovered topics in a set of documents and identify the discovered topics in real time.

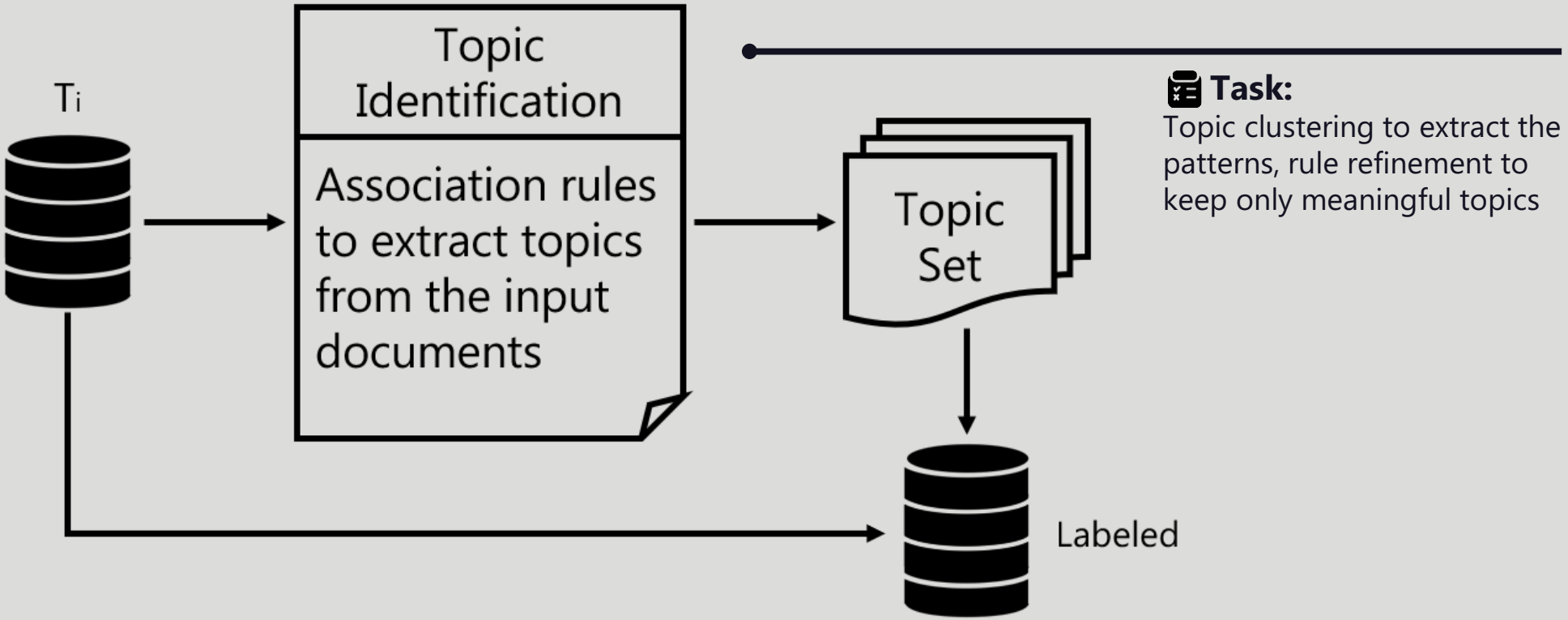
Task:
Pre-process pipeline to normalize documents, applying the previously seen



Task:
Database segmentation:
Identifier *(T_0-T_j)*
Modeler *(T_{j+1}-T_k)*
Validation *(T_{k+1}-T_n)*

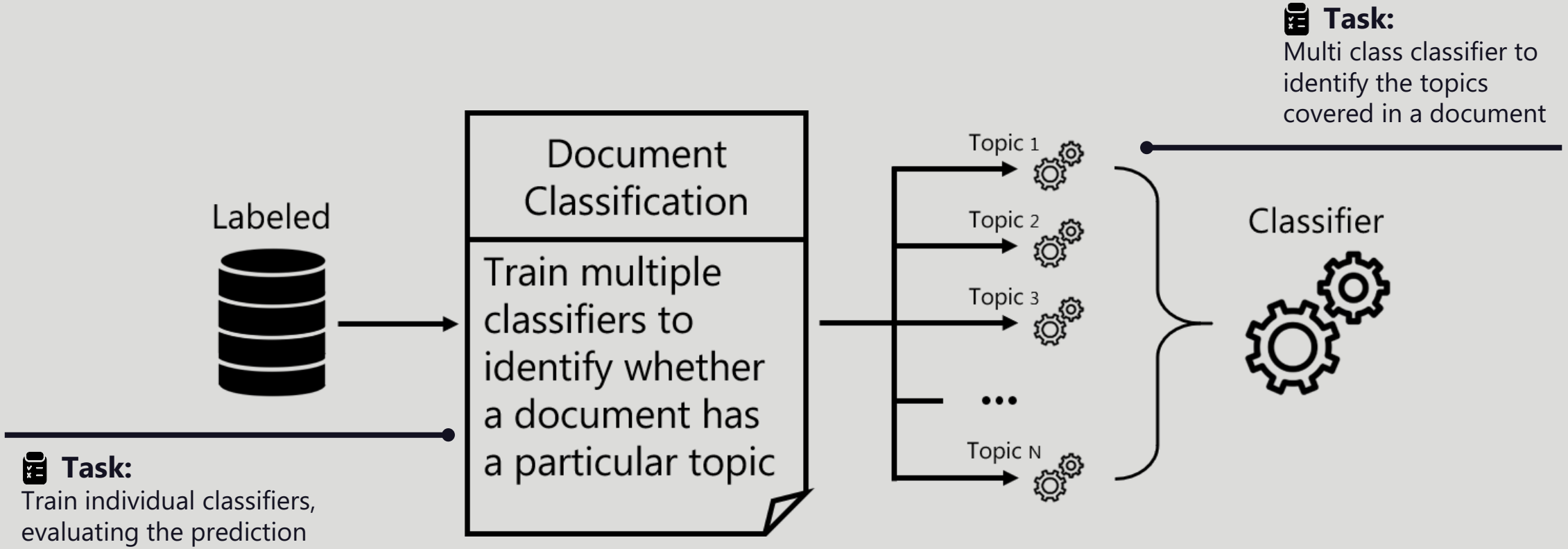
MATERIALS AND METHODS | Topic Identification

As discussed earlier, we want to build models capable of make predictions regarding the evolution of discovered topics in a set of documents and identify the discovered topics in real time.



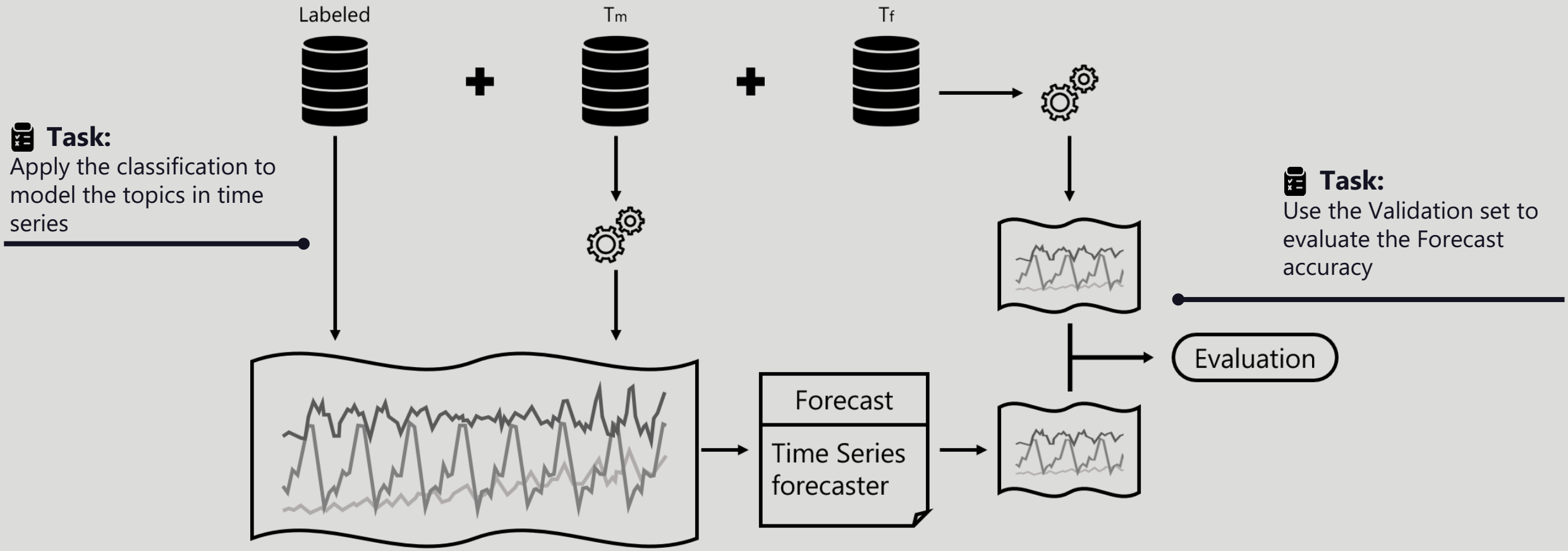
MATERIALS AND METHODS | Document Classification

As discussed earlier, we want to build models capable of make predictions regarding the evolution of discovered topics in a set of documents and identify the discovered topics in real time.



MATERIALS AND METHODS | Forecast Evaluation

As discussed earlier, we want to build models capable of make predictions regarding the evolution of discovered topics in a set of documents and identify the discovered topics in real time.



CONTENT

ROADMAP

In view of the problem's complexity, we can elaborate a schedule with the proposed tasks in the previously. The table above show the tasks over the remains months until the end of this work.

Sprint	Start Date	End Date	Duration	Task
#1	August 3	August 16	14 days	- Choose a database - Pre process the database
#2	August 17	September 6	21 days	- Topic Identification
#3	September 7	September 27	21 days	- Document Classification
#4	September 28	October 18	21 days	- Time Series Forecast
#5	October 19	November 8	21 days	- Test and fix bugs

Obrigado!