Natural Language Processing for Trend Forecasting

Heládio Sampaio Lopes

Computer Engineering (ITA 2020)

CONTENT

CONTENT

INTRODUCTION

Natural Language Processing can perform a lot of tasks, such as identifying different topics for a set of documents, classifying texts on predefined subjects, and beyond that extract the sentiment to know what people are saying about something.

Motivation

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**.

Objectives

Explore Natural Language Processing techniques to propose a framework to **modeling in real-time the topics**' evolution, and evaluate these models ability to make **predictions** about **future trends**.

CONTENT

LITERATURE REVIEW | Text Processing Techniques

The key task to several machine learning problems consists in make a good data representation before applying any model. A clean data set can allow a model to increase its performance in the learning process, making a better identification in the patterns present in the variables.

Normalization

Tokenization

Bag of Words

TF-IDF

LITERATURE REVIEW | Text Processing Techniques

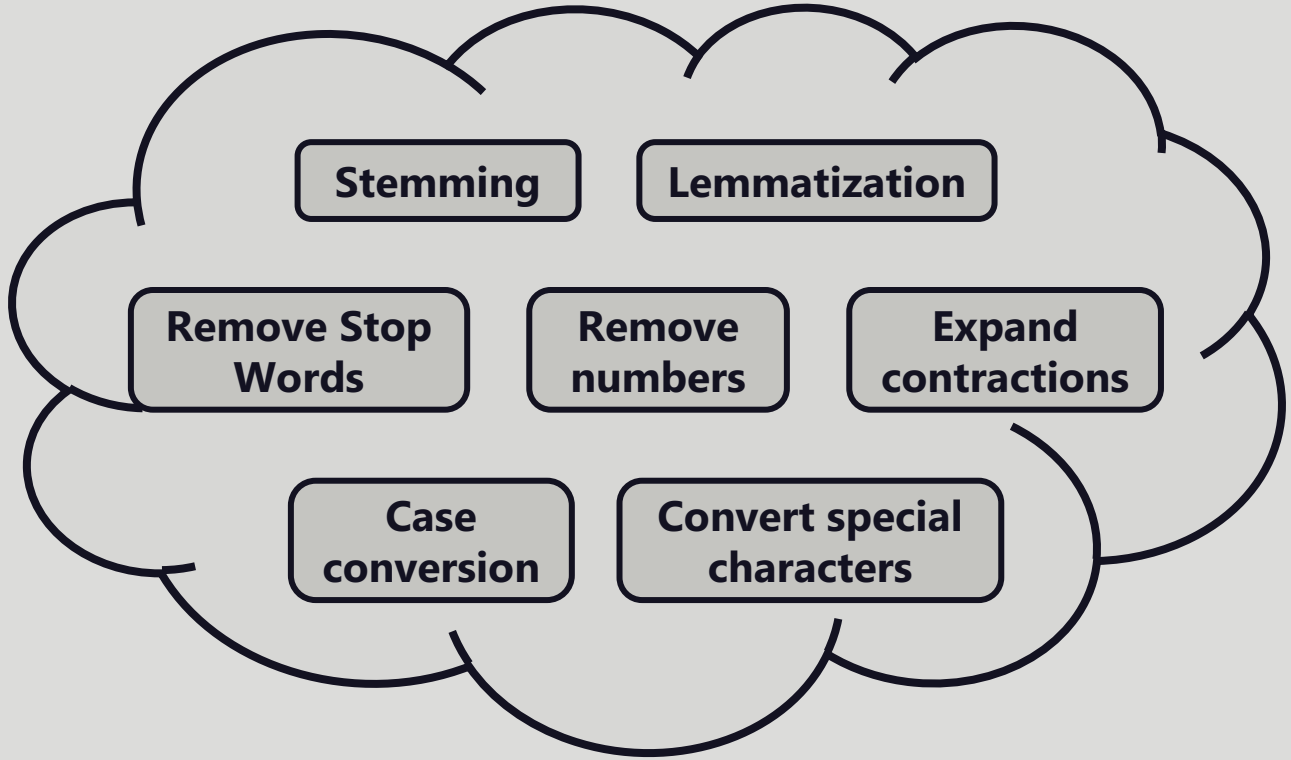
The key task to several machine learning problems consists in make a good data representation before applying any model. A clean data set can allow a model to increase its performance in the learning process, making a better identification in the patterns present in the variables.

Normalization

Tokenization

Bag of Words

TF-IDF



LITERATURE REVIEW | Text Processing Techniques

The key task to several machine learning problems consists in make a good data representation before applying any model. A clean data set can allow a model to increase its performance in the learning process, making a better identification in the patterns present in the variables.

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

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

LITERATURE REVIEW | Text Processing Techniques

The key task to several machine learning problems consists in make a good data representation before applying any model. A clean data set can allow a model to increase its performance in the learning process, making a better identification in the patterns present in the variables.

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
Doc 1:	1	1	1	1	1	1	0	0	0	0	0	0
Doc 2:	1	1	0	0	1	0	1	1	1	0	0	0
Doc 3:	1	1	1	0	1	0	0	0	0	1	1	1

LITERATURE REVIEW | Text Processing Techniques

The key task to several machine learning problems consists in make a good data representation before applying any model. A clean data set can allow a model to increase its performance in the learning process, making a better identification in the patterns present in the variables.

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_K
Doc 1:					
Doc 2:					
Doc 3:					

Variables

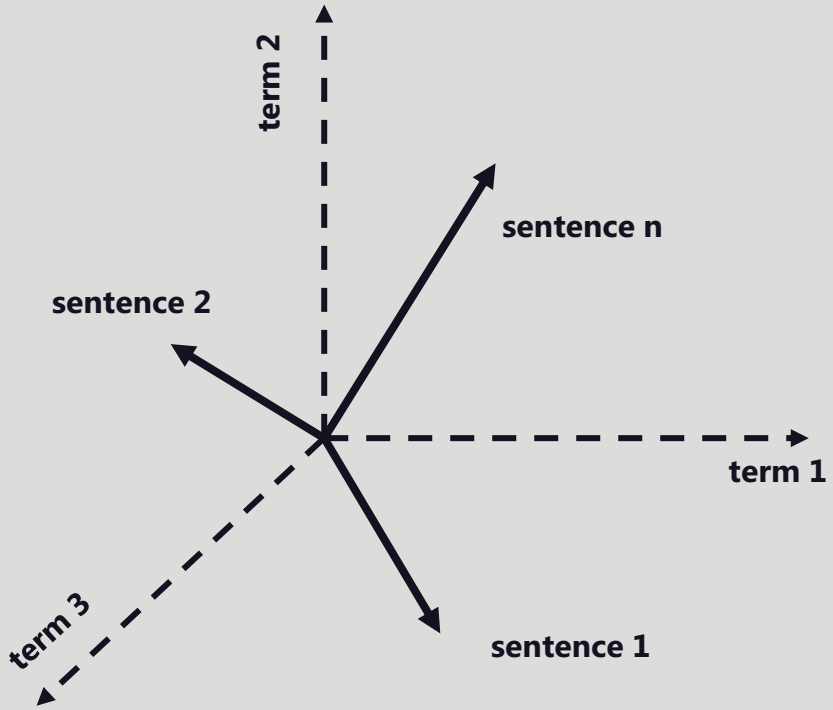
- $w_{i,j}$: Word Weight
- $\mathrm{tf}_{i,j}$: Term Frequency
- df_i : Document Frequency
- N : Size of Document Set

LITERATURE REVIEW | Word Embedding

The Word Embedding is a technique to represent words in vectors capable of capture the words context in a document. It is also able to smooth the high dimensionality effect by using much more compact vector to represent the words.

Traditional approach

- Word's vector space
- Problems:
  - Sparse representation
  - High dimensionality
  - Distinct vectors for words with the same meaning
  - Have a good day (x) Have a nice day

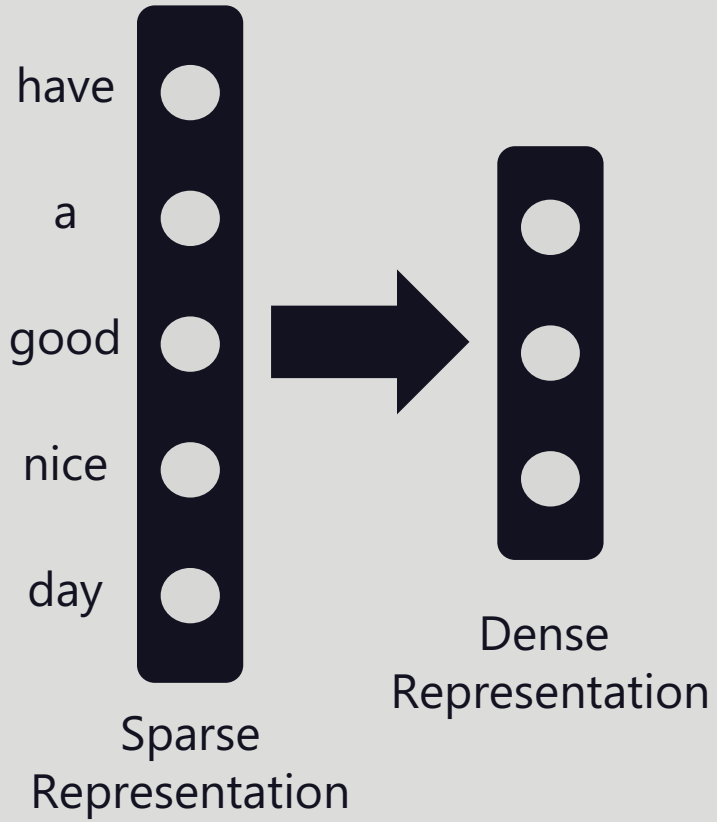


LITERATURE REVIEW | Word Embedding

The Word Embedding is a technique to represent words in vectors capable of capture the words context in a document. It is also able to smooth the high dimensionality effect by using much more compact vector to represent the words.

Word Embedding

- Dense representation
- Decreases dimensionality
- Similar vectors for words with the same meaning
- Frameworks:
  - Word2Vec
  - Glove
  - Fast Text



LITERATURE REVIEW | Topic Modeling

In NLP, topic modeling is frequently used text-mining tool to identify hidden patterns, called “topics”, in a collection of documents. Latent Dirichlet Allocation, LDA for short, is a statistical model which uses a Dirichlet distribution to model both the topics and the words, after finding the patters it is human work identify which topics make sense.

Food **Technology** **Sports**

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.

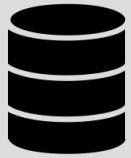
Trend Forecast

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.



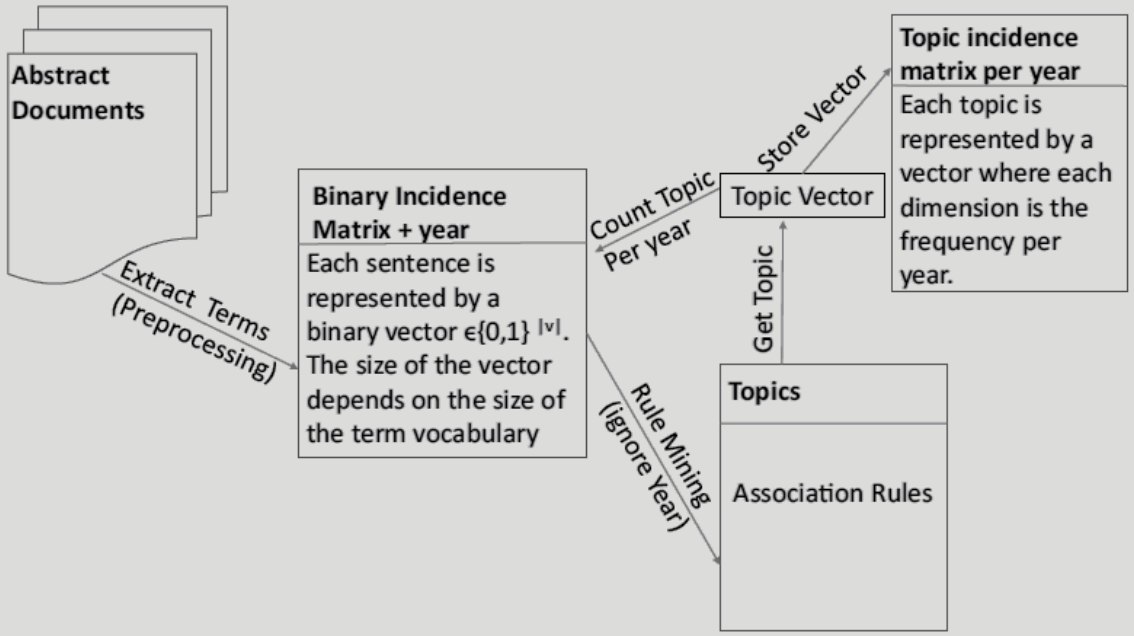
Data set containing titles and abstracts over more than six thousand academic papers distributed between 2002 and 2010, data acquired by Tang *et al.* (2008).

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

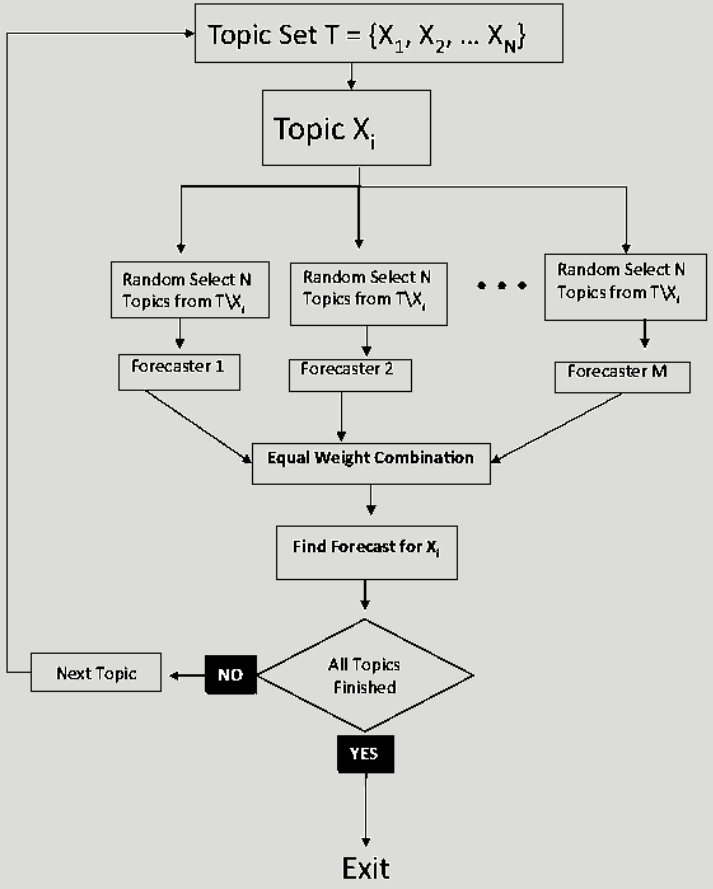
RELATED WORKS

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

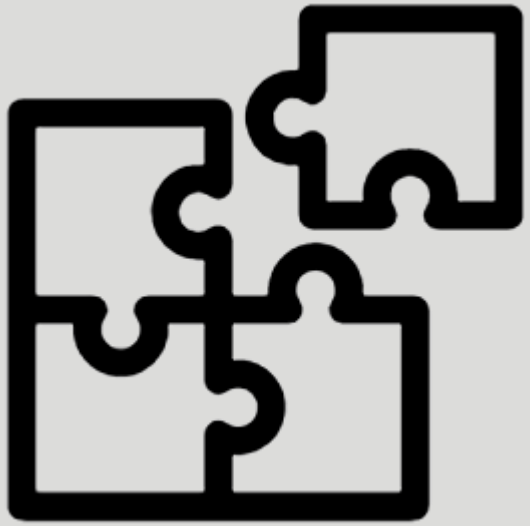
Hurtado *et al* (2016) use topic modeling to inspect research publications, patents, and technical reports aiming to model the evolution of the direction of research and forecast the near future trends in IT industry.



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

RELATED WORKS | Gap

Topic discovery and trend forecast were subjects widely explored in the literature. With that in mind, we wish, in this work, to reproduce these techniques. However, in addition to what has been presented we want to be able to explore some modifications.



Real-time system that will keep receiving news in a continuous process.

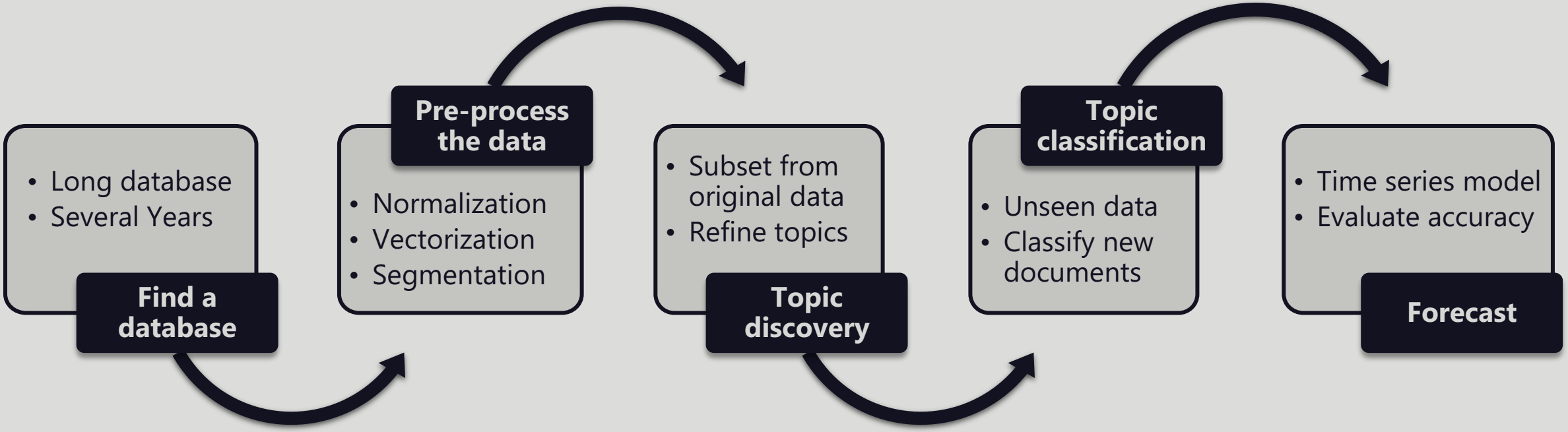
Redo the discovery process will demand an expensive computational cost.

New topic classification-based system will be proposed in order fill this gap.

CONTENT

MATERIALS AND METHODS | Objectives

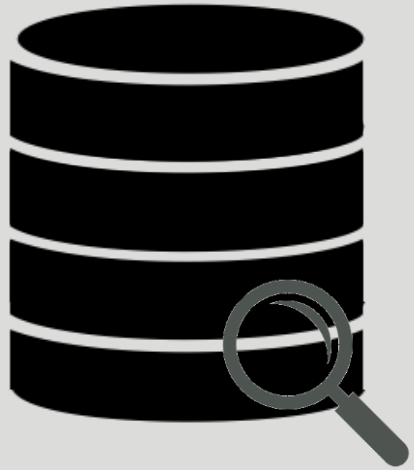
As discussed earlier, we want to build models capable of make predictions regarding the evolution of discovered topics in a set of documents and identify the discovered topics in real time.



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.

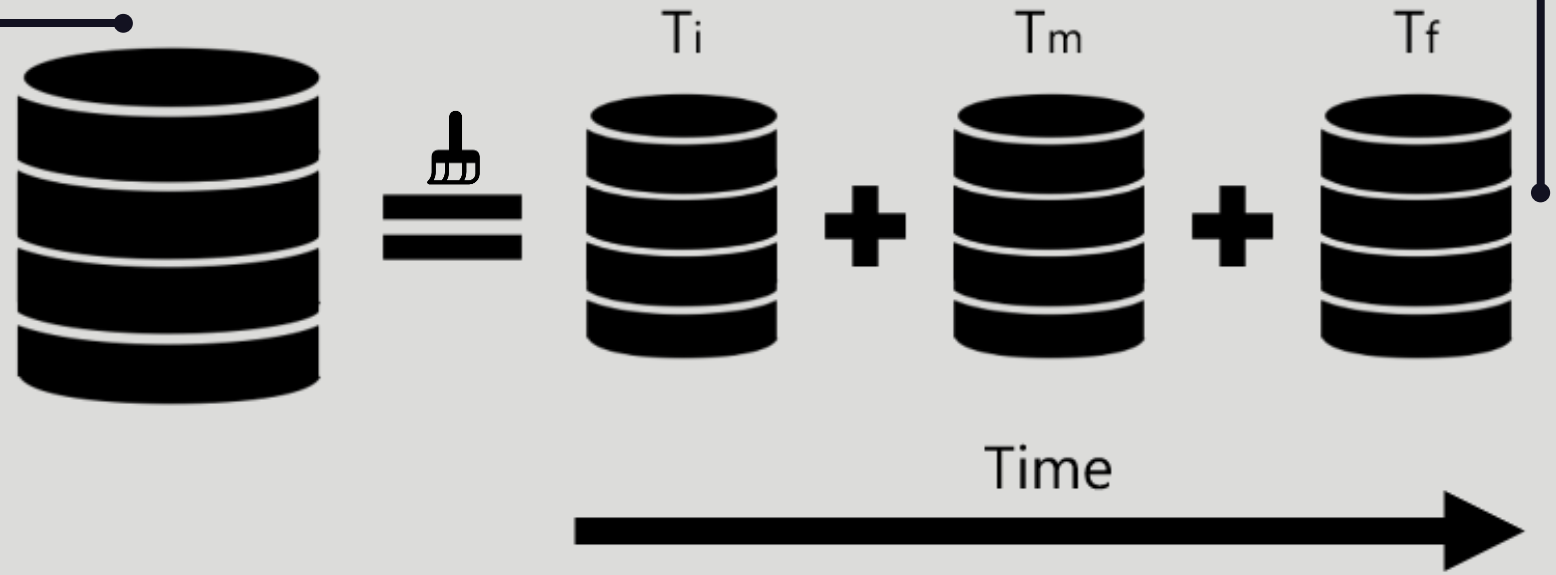
- Wikipedia Daily News
- Newspapers Articles
- Academic Papers
- Social Media - Reddit



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.

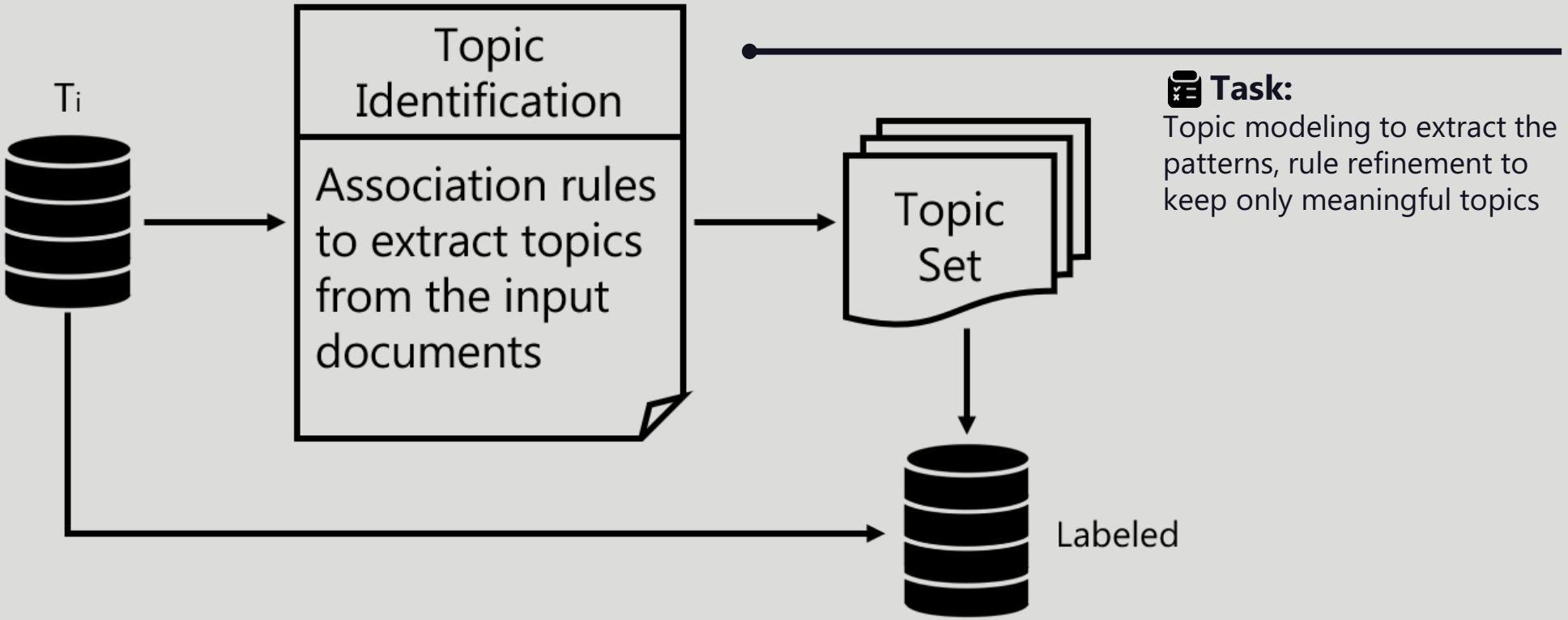
Task:
Pre-process pipeline to normalize documents, applying the previously seen



Task:
Database segmentation:
Identifier $(t_0\text{-}t_j)$
Modeler $(t_{j+1}\text{-}t_k)$
Validation $(t_{k+1}\text{-}t_n)$

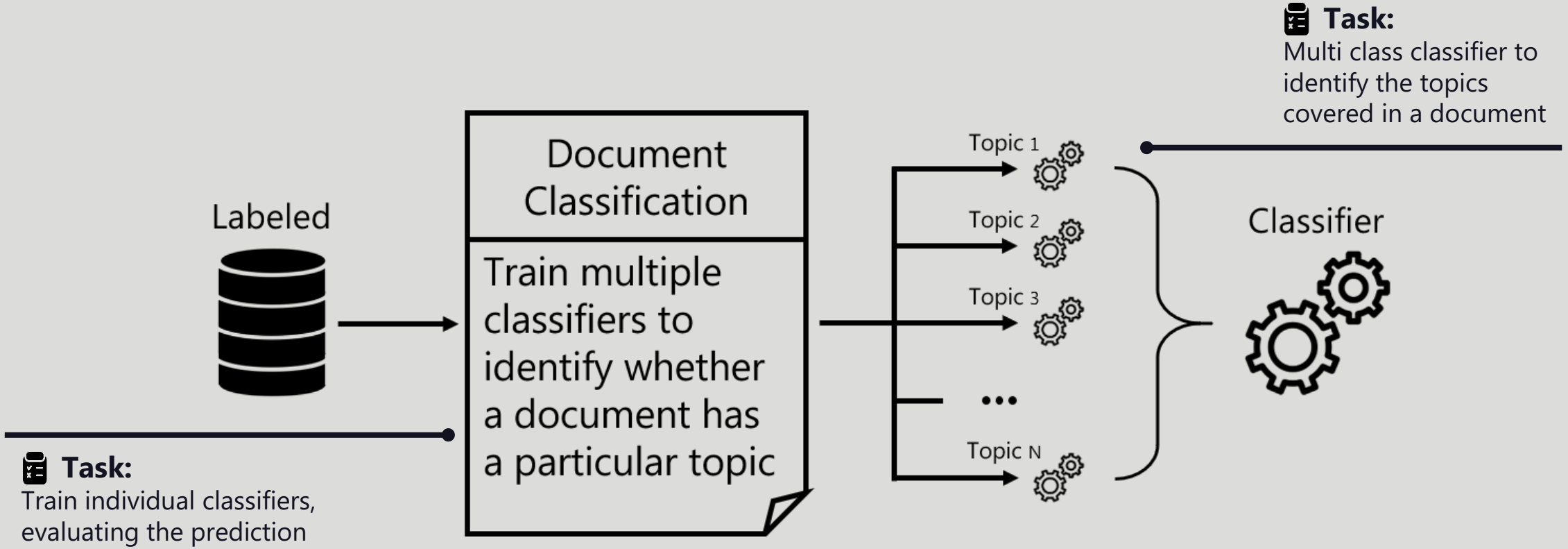
MATERIALS AND METHODS | Topic Identification

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.



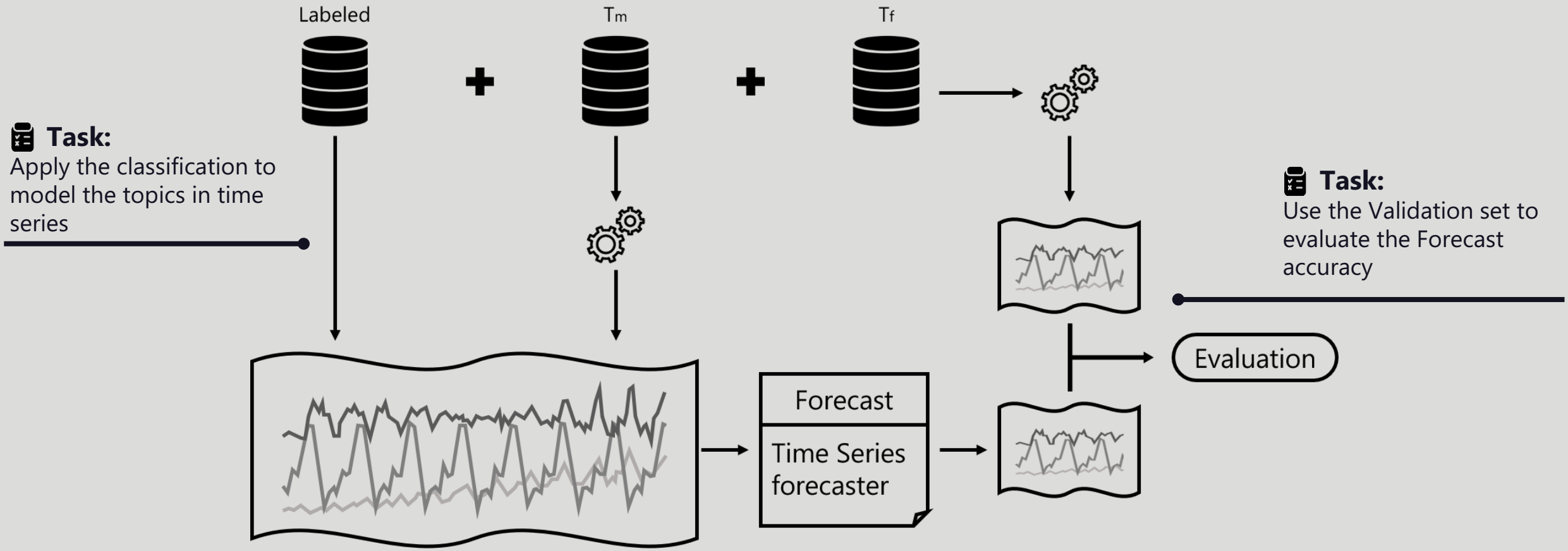
MATERIALS AND METHODS | Document Classification

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	August 20	14 days	- Topic Identification
#3	August 31	September 20	21 days	- Evaluate Identification - Generate Identification Results
#4	September 21	October 4	14 days	- Document Classification
#5	October 5	October 25	21 days	- Evaluate Classification - Generate Classification Results
#6	October 26	November 8	14 days	- Test and fix bugs

Thank you!