EXTRACTIVE SUMMARIZATION FOR EXPLAINABLE SENTIMENT ANALYSIS USING TRANSFORMERS {EXS4EXSA}

X-SENTIMENT

6TH INTERNATIONAL WORKSHOP ON
EXPLAINABLE SENTIMENT MINING AND EMOTION DETECTION

EUROPEAN SEMANTIC WEB CONFERENCE 2021 - ESWC 2021

L. BACCO - A. CIMINO - F. DELL'ORLETTA - M. MERONE

JUNE 7, 2021

ABSTRACT

R&D QUESTION

Sentiment Analysis: since more and more content is shared by people on the web, automated SA tools have been employed in several tasks, such as

Social media monitoring

Customer care services

Market research

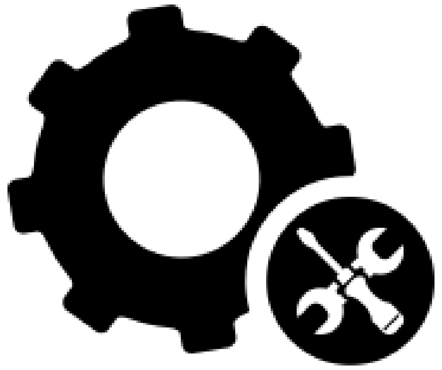
but often lack transparency...

ABSTRACT

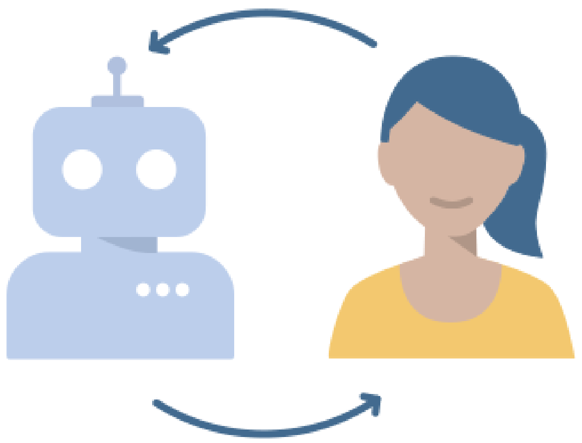
R&D QUESTION

Explainability: End users and companies, developers and research communities, even governative organizations (*Articles 13-15, 22 of the EU GDPR*) are demanding for e**X**plainable **A**rtificial **I**ntelligence systems **[1]**.
XAI models should provide human-interpretable explanation of their decisions to

Increase trust and reliability of the user



Exploit insights from the models to improve the building pipelines



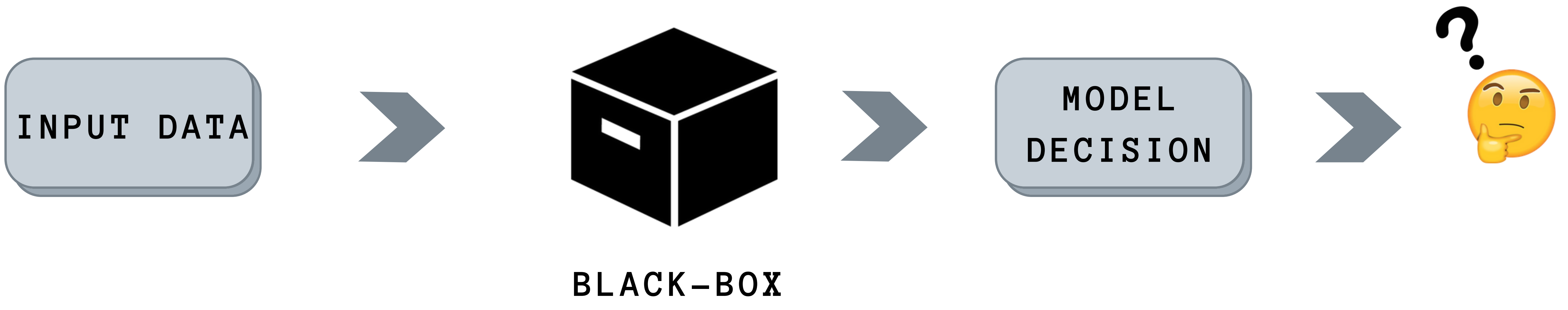
Ensure rights of explanation and to opt-out model decisions

without affecting the models performance

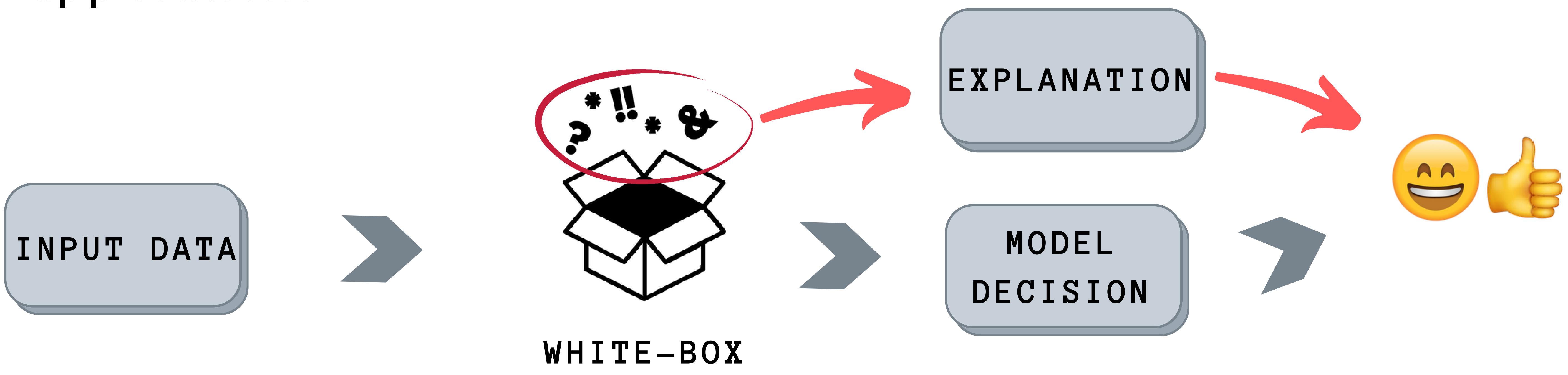
EXS4EXSA

ABSTRACT
EXPLAINABILITY

- Past AI applications



- Future AI applications



EXS4EXSA

ABSTRACT

MAIN CONTRIBUTIONS

- **A new approach** to explain document classification tasks as sentiment analysis, by providing extractive summaries as the explanation of the model decision

- **Exploring the use of attention weights** of a hierarchical transformer architecture as a base to achieve extractive summaries explanation

- **A new annotated dataset*** for the evaluation of the extractive summaries as an explanation of a sentiment analysis task.

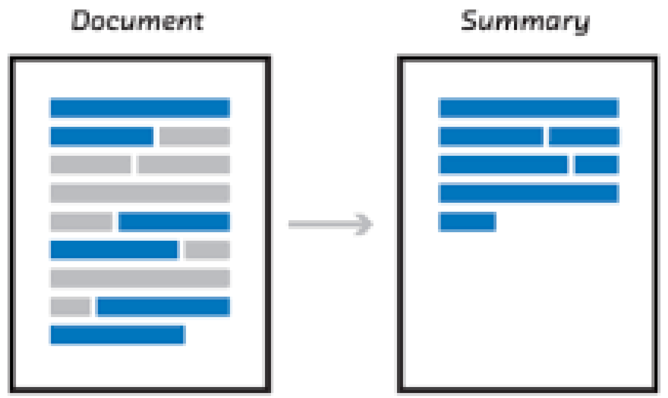
- **Two different proposed models**, both based on transformer architectures, analyzed in terms of the performance in both the classification and explanation tasks.

*www.github.com/lbacco/ExS4ExSA

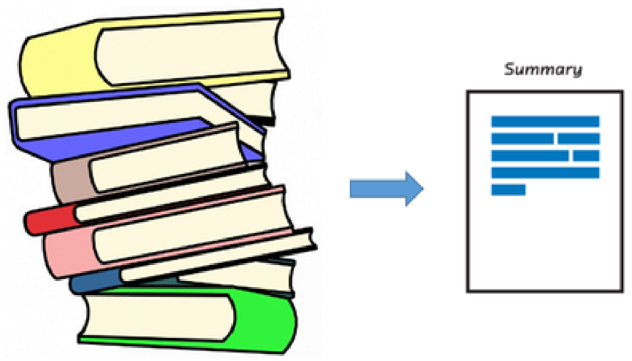
CONCEPTS

SUMMARIZATION [2]

Single-Document

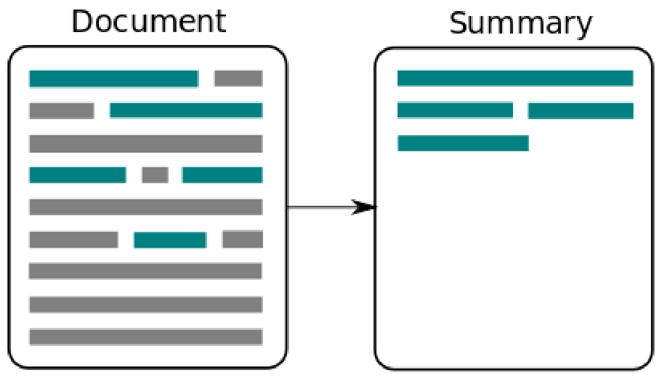


Vs.



Multi-Documents

Extractive:
- Easier task
- Redudancy and information lost



Vs.

Abstractive:
- Very complex task
- Helps reducing the issues

Supervised

Vs.

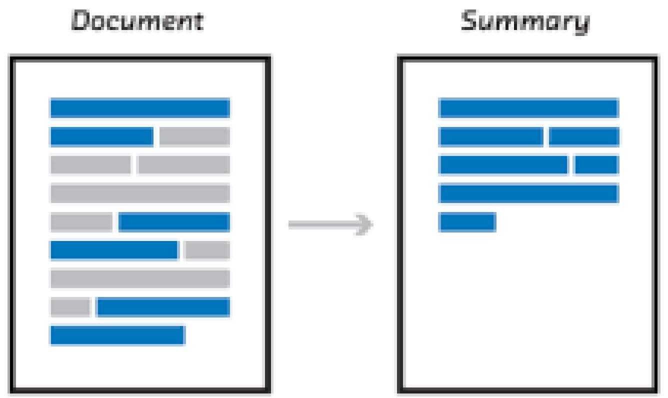
Unsupervised (or self-supervised)

EXS4EXSA

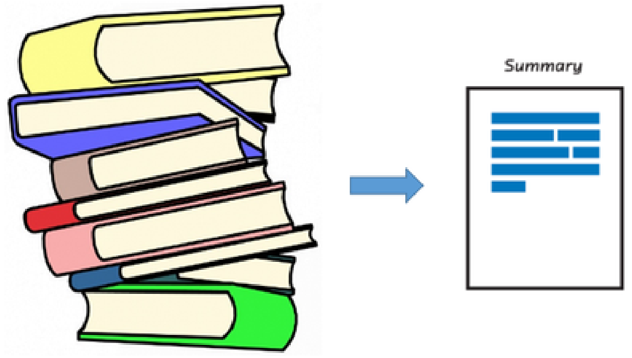
CONCEPTS

SUMMARIZATION [2]

Single-Document

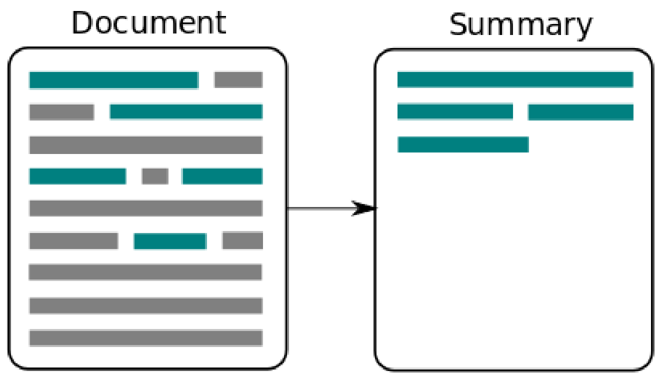


Vs.



Multi-Documents

Extractive:
- Easier task
- Redudancy and information lost



Vs.

Abstractive:
- Very complex task
- Helps reducing the issues

Supervised

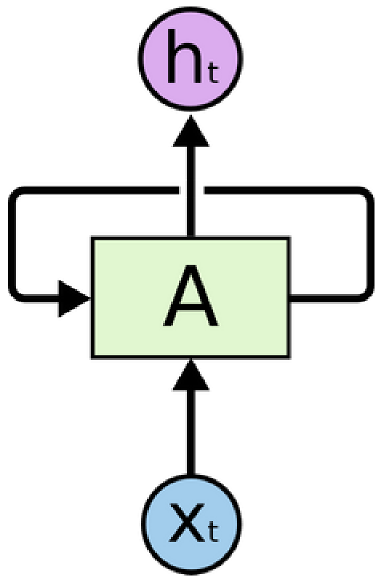
Vs.

Unsupervised (or self-supervised)

EXS4EXSA

CONCEPTS

TRANSFORMER VS RNN



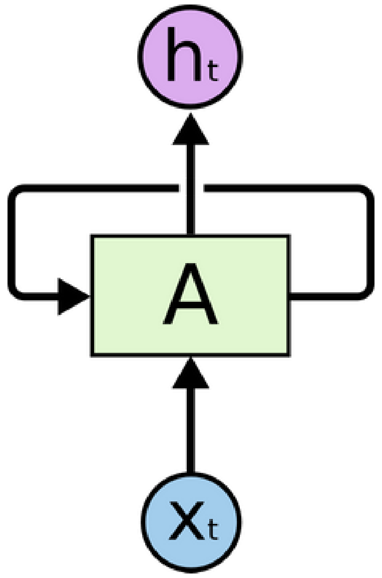
RNNs, aka Recurrent Neural Networks.
Years ago, they were the preferred solution to capture context dependencies in text:

- Relatively small # of parameters
- Intrinsically sequential
- Relatively small context dependency (due to the Exploding/Vanishing Gradients issues)

EXS4EXSA

CONCEPTS

TRANSFORMER VS RNN



RNNs, aka Recurrent Neural Networks.
Years ago, they were the preferred solution to capture context dependencies in text:

- Relatively small # of parameters
- Intrinsically sequential
- Relatively small context dependency (due to the Exploding/Vanishing Gradients issues)

Transformers [3,4]. Pretty obiquitous in the recent years literature

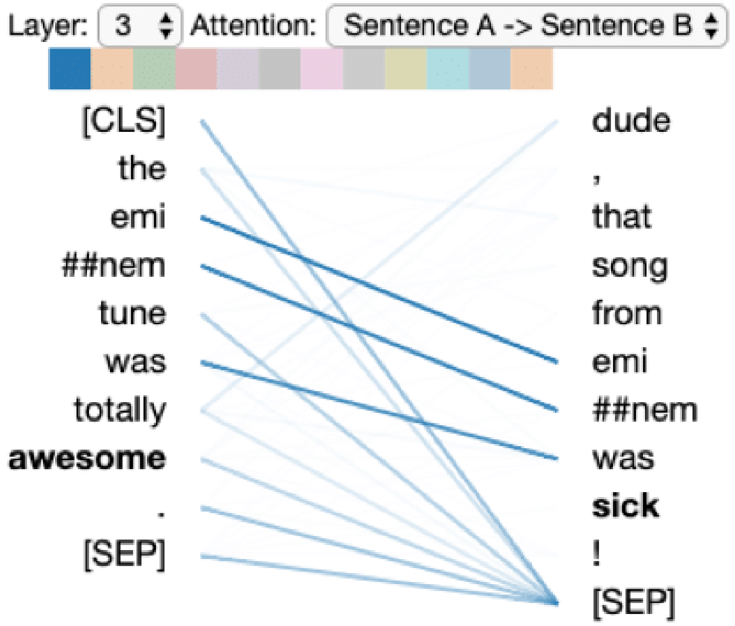
- (Multi-head self-)attention mechanisms
-
-

- Highly parallelizable
- Longer-term context dependency

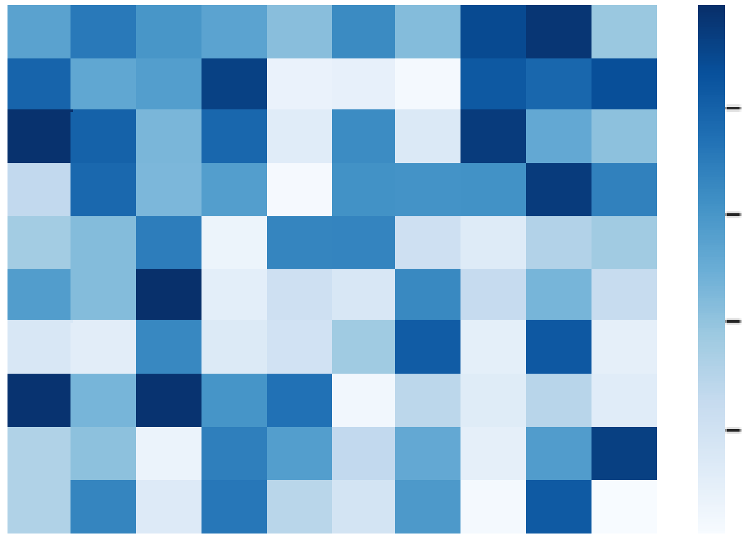
EXS4EXSA

CONCEPTS

ATTENTION AS EXPLANATION



BertViz [5]



Heatmaps

Tools to **visualize attention** weights inside Transformer models, already used in literature to:

- find out **language properties** from the self-attention heads **[5]**
- **highlight** most important **n-grams** in the text **[6]**

Transformers [3,4]. Pretty obiquitous in the recent years literature

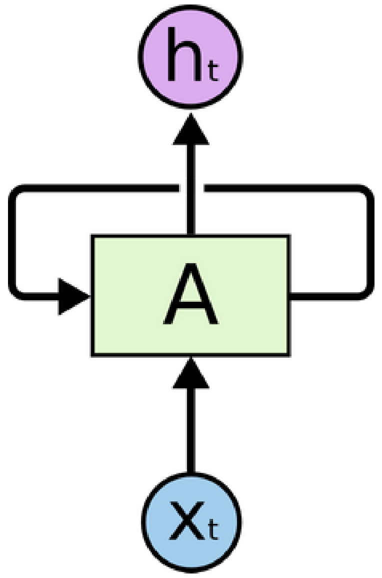
- (Multi-head self-)attention mechanisms
-
-

- Highly parallelizable
- Longer-term context dependency

EXS4EXSA

CONCEPTS

TRANSFORMER VS RNN



RNNs, aka Recurrent Neural Networks.
Years ago, they were the preferred solution to capture context dependencies in text:

- Relatively small # of parameters
- Intrinsically sequential
- Relatively small context dependency (due to the Exploding/Vanishing Gradients issues)

Transformers [3,4]. Pretty obiquitous in the recent years literature

- (Multi-head self-)attention mechanisms
- Very large # of parameters
-

- Intrinsically limited to "short" texts

EXS4EXSA

CONCEPTS

HIERARCHY IN TRANSFORMERS

Self-attention mechanisms require high computational and memory resources.
There are two approaches to deal with this issue:

- Truncation (simplest, leads to lost of information)
- Hierarchy (already used in document classification **[7]** and summarization tasks **[8]**)

...some chunk of text... → TRANSFORMER → New representation → Some model → Decision

Working at token level

Working at chunk level

Transformers [3,4]. Pretty obiquitous in the recent years literature

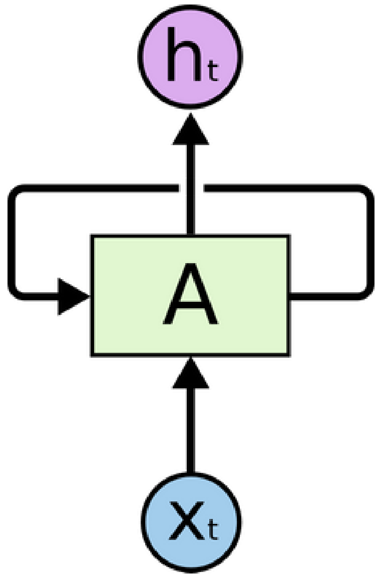
- (Multi-head self-)attention mechanisms
- Very large # of parameters
-

- Intrinsically limited to "short" texts

EXS4EXSA

CONCEPTS

TRANSFORMER VS RNN



RNNs, aka Recurrent Neural Networks.
Years ago, they were the preferred solution to capture context dependencies in text:

- Relatively small # of parameters
- Intrinsically sequential
- Relatively small context dependency (due to the Exploding/Vanishing Gradients issues)

Transformers [3,4]. Pretty obiquitous in the recent years literature

- (Multi-head self-)attention mechanisms
- Very large # of parameters
- SOA performance in several NLP tasks ✓

- Highly parallelizable
- Longer-term context dependency

- Intrinsically limited to "short" texts

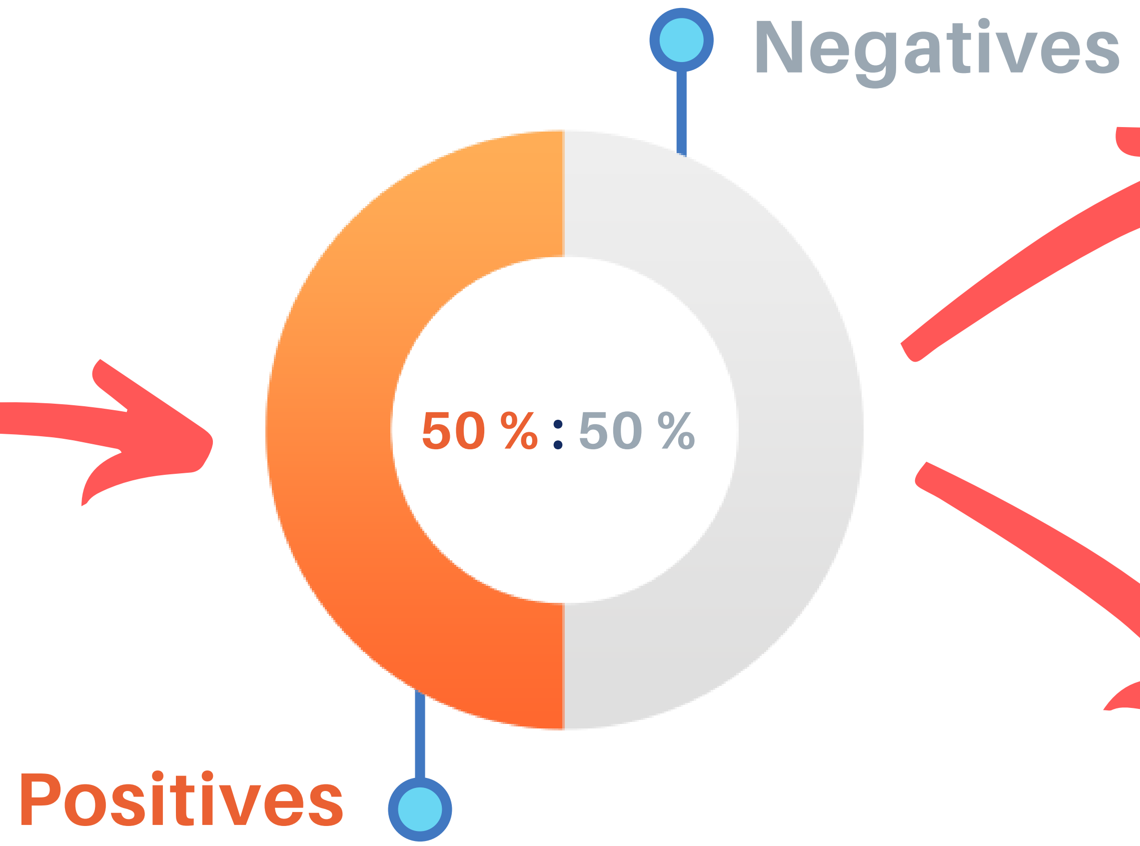
EXS4EXSA

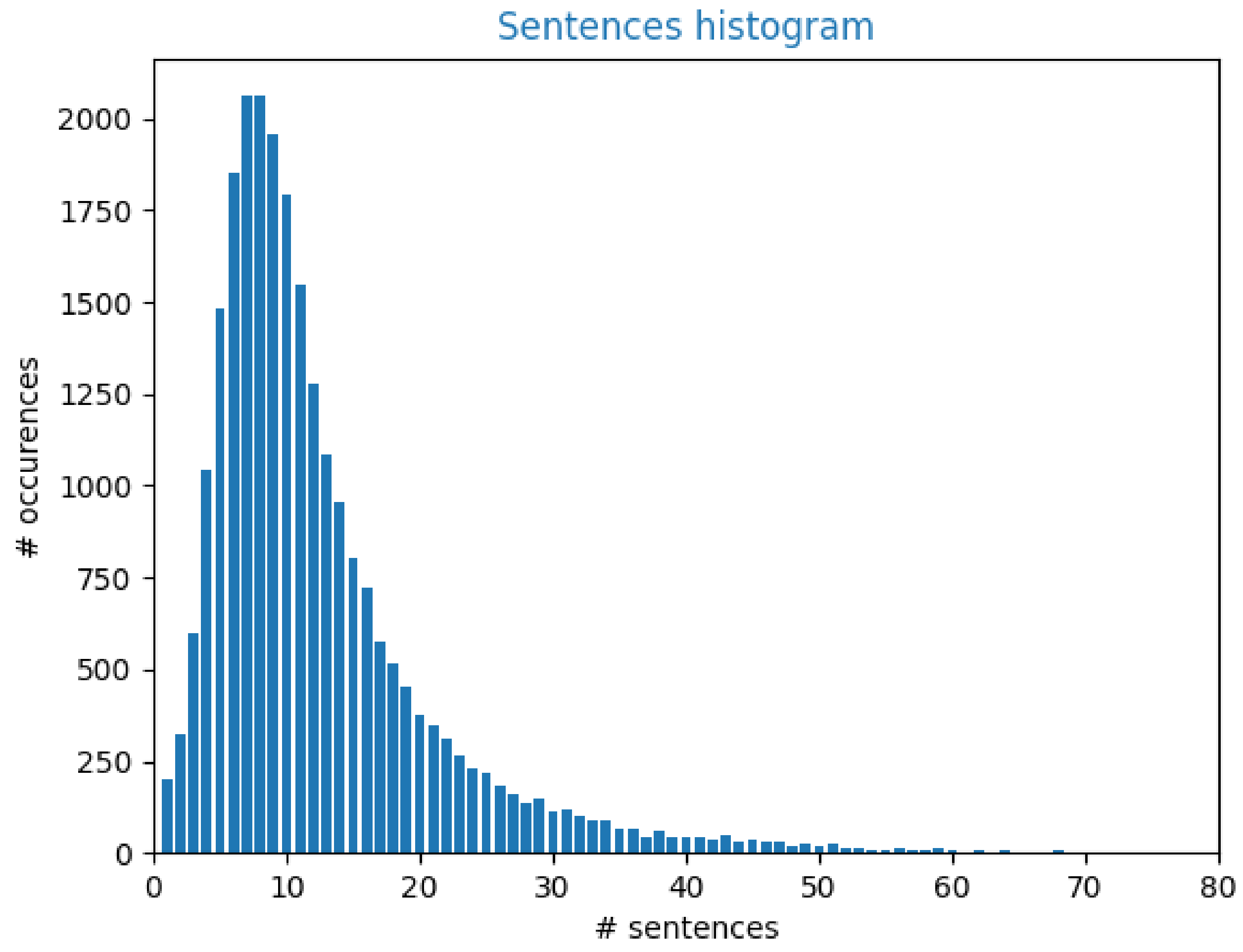
DATASET

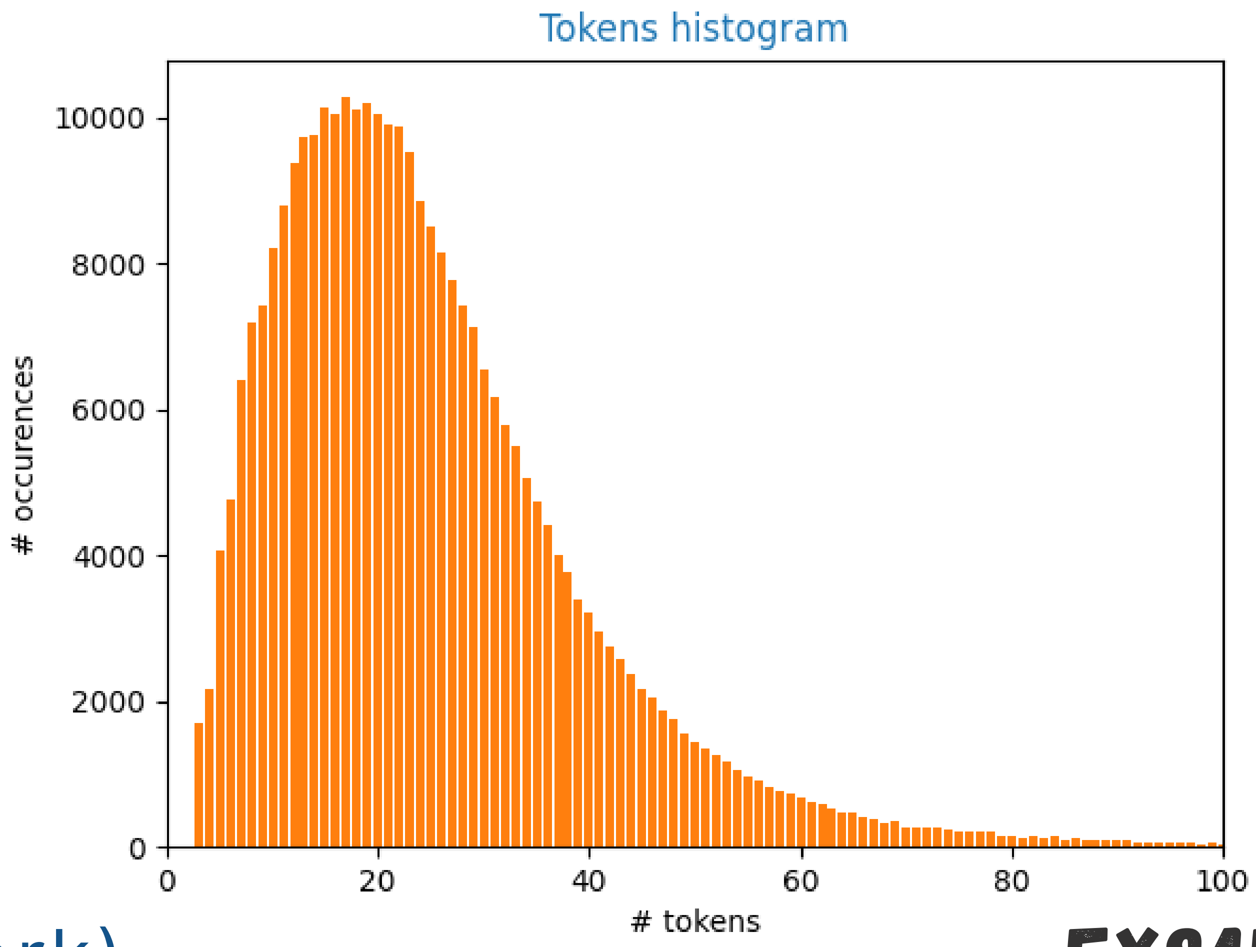
DATA ANALYSIS

[9]

IMDb







- **Training set:** 25k samples
- **Test set:** 25k samples
- **Unlabeled set:** 50k samples (not used in this work)

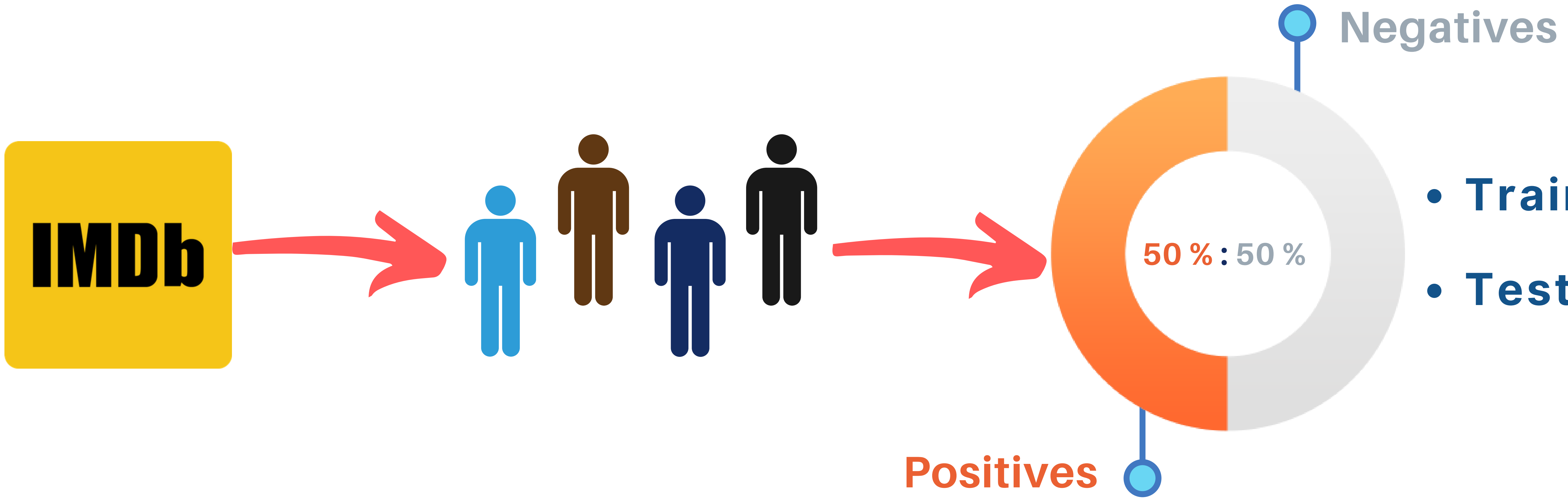
EXS4EXSA

DATASET

DATA ANNOTATION

4 **independent annotators** selects the 3 **most important sentences** in each document, while **taking into account the sentiment** of the document.

[9]



- **Training set:** 50 samples
- **Test set:** 100 samples

The quality of their annotations was evaluated through the **Krippendorff's Alpha*** inter-annotators index **[10]**

α

- **Training set:** 0.47
- **Test set:** 0.61

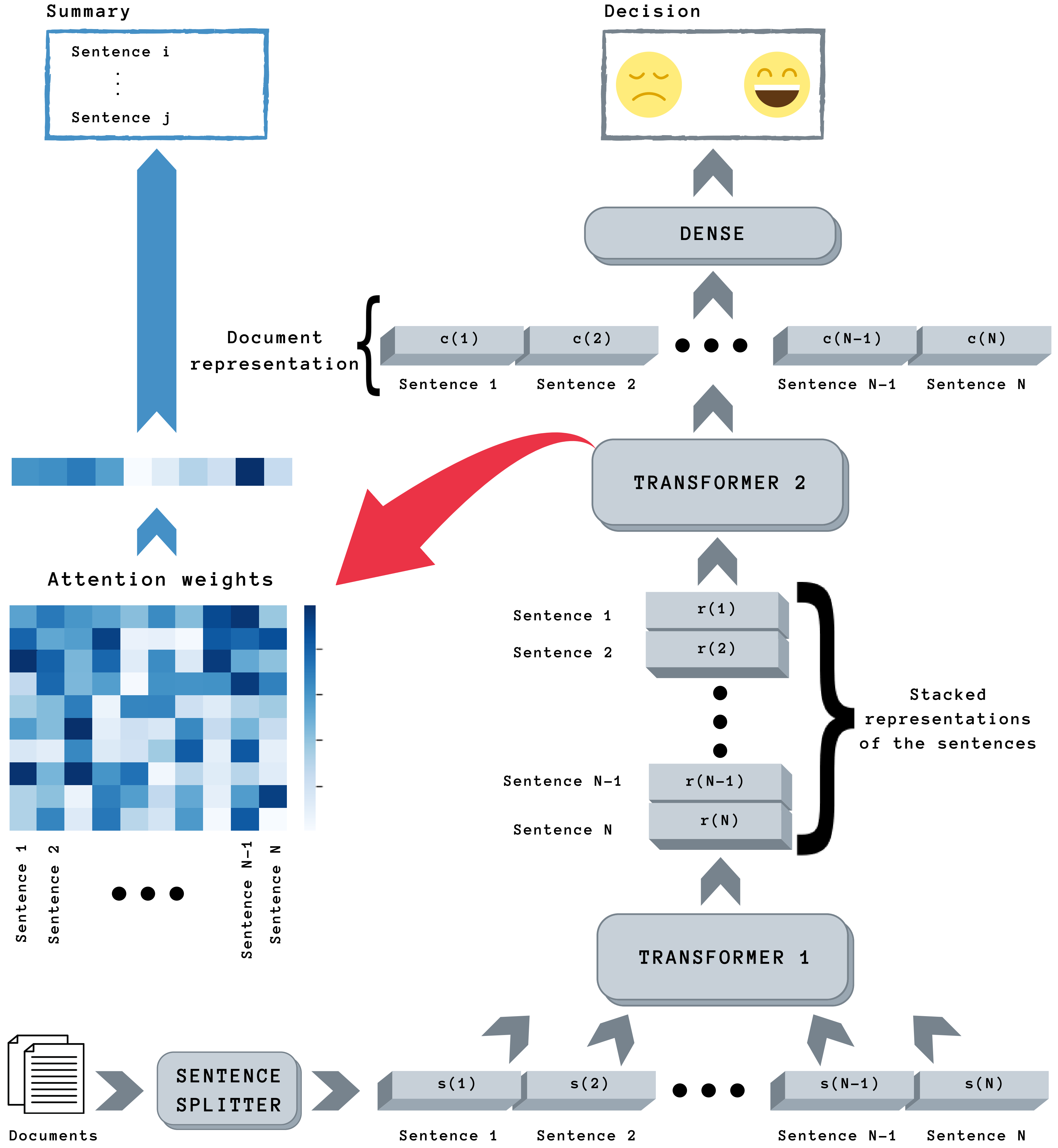
*www.github.com/lbacco/ExS4ExSA

EXS4EXSA

EXPLAINABLE MODELS

HIERARCHICAL TRANSFORMER

- **T1:** the first transformer
- **s(i):** input of **T1**, i.e the i-th sentence
- **r(i):** output of **T1**, i.e. the new representation of the i-th sentence
- **T2:** the second transformer
- **Merging layer:** implements a merging strategy (average, concatenation etc.)
- **d:** output of the merging layer, i.e the document representation
- **Dense layer:** outputs the predicted sentiment based on **d**
- **Attention weights:** extracted from **T2** and **ranked** to build the summary

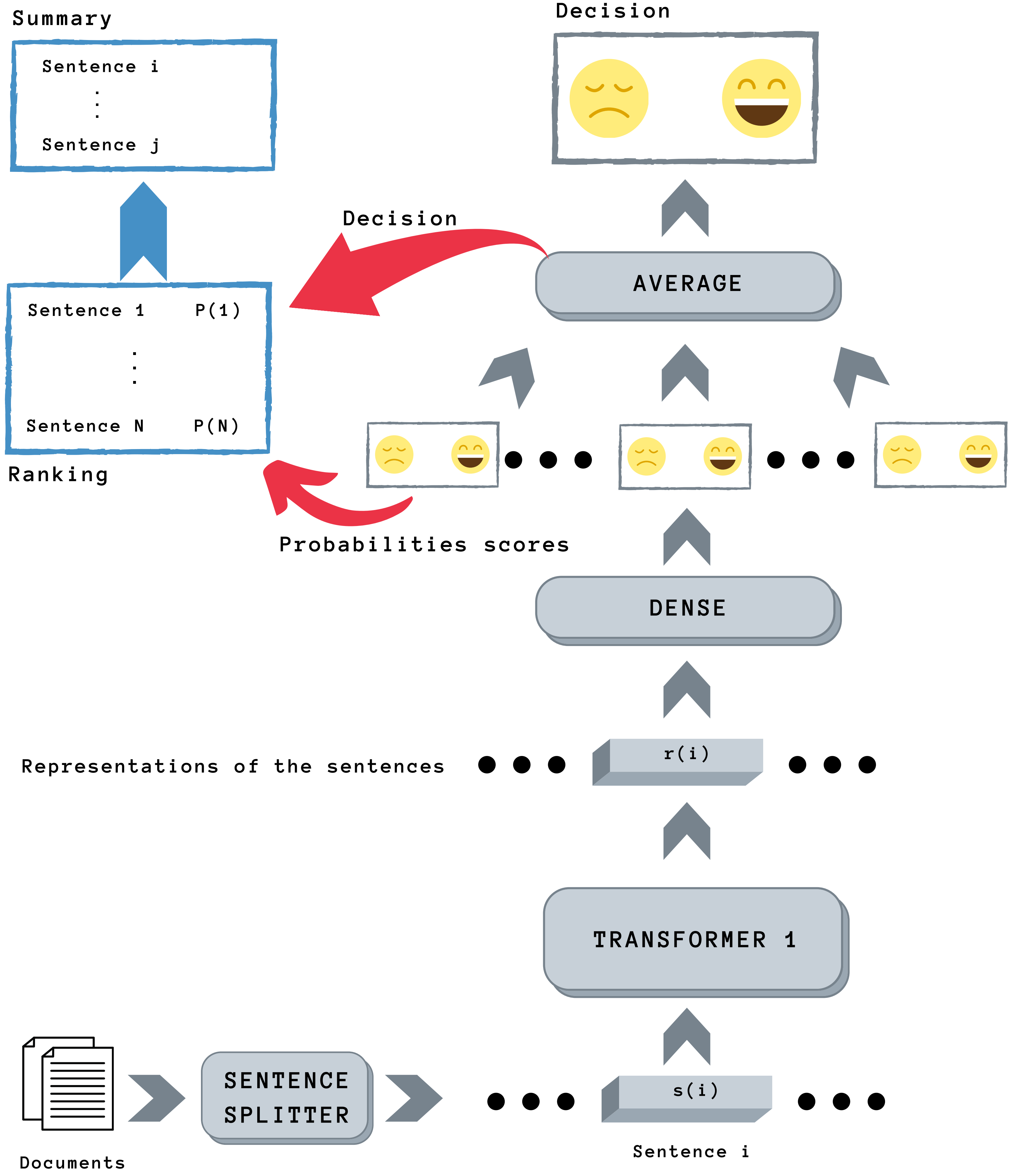


EXS4EXSA

EXPLAINABLE MODELS

SENTENCE CLASSIFICATION COMBINER

- **T1:** the first transformer
- **s(i):** input of **T1**, i.e the i-th sentence
- **r(i):** output of **T1**, i.e. the new representation of the i-th sentence
- **Dense layer:** outputs two probabilities scores for the sentiment of the i-th sentence
- **Average layer:** implements the sentiment-specific average of the probabilities scores; the predicted document sentiment is the sentiment with the greatest average result
- **Probabilities scores of the winner sentiment:** ranked to build the summary



EXS4EXSA

EXPLAINABLE MODELS

EXPERIMENTS

Main hyperparameters:

- **N:** the maximum number of sentences for each document was set equal to **15**, in order to cover the **75%** of all the documents without truncation while limiting the computational effort. Empty sentences were added to the documents with less that 15 sentences for the hierarchical model.
- **t:** the maximum number of tokens for each sentence was set equal to **32**, in order to cover the **75%** of all the sentences whitout truncation while limiting the computational effort. In both the models, sentences with less than 32 tokens were right-padded and padded regions were masked.
- **T1:** the first transformer is the **RoBERTa Transformer**, an optimized version of BERT. The same T1 has been chosen for both the models in order to allow a fair comparison.
- **Merging strategies:**
  - Concatenation
  - Average
  - Masked Average
  - BiLSTM
- **T2:** the second transformer in the hierarchical model consists of:
  - 2 layers
  - 1 attention head per layer

  These parameters were chosen to ease the attention weigths-based explainability considerations.

RESULTS

SENTIMENT ANALYSIS

- **ExHiT: Ex**plainable **Hi**erarchical **T**ransformer
- **SCC: S**entence **C**lassification **C**ombiner

Model	Merging strategy	Accuracy (%)	Precision (%)		Recall (%)	
			Neg	Pos	Neg	Pos
ExHiT	Concatenation	92.59	90.97	**94.34**	**94.56**	90.62
	Average	92.35	92.18	92.51	92.54	92.15
	Masked Average	92.77	92.07	93.49	93.60	91.94
	BiLSTM	92.34	90.97	93.80	94.01	90.67
SCC	-	**93.51**	**95.42**	91.75	91.40	**95.62**

- The **SCC** model slightly outperforms the **ExHiT** model
- Changing the **merging strategies** in the **ExHiT** architecture does not have a great impact on the performances
- **SCC** resulted particularly good for the precision for the negative class and the recall for the positive one, while achieving the worst performances for their counterpart metrics

RESULTS

EXPLAINABILITY

- **ExHiT: Ex**plainable **Hi**erarchical **T**ransformer
- **SCC: S**entence **C**lassification **C**ombiner

Model	Merging strategy	Agreement at least 1 Precision (%)		Agreement at least 2 Precision (%)		Agreement at least 3 Precision (%)	
		test	train	test	train	test	train
ExHiT	Concatenation	53.82%	55.88%[a]	49.15%	45.00%	46.63%	46.45%
	Average	58.04%	57.82%	50.42%	45.92%[1]	45.29%	41.84%
	Masked Average	53.15%[a]	55.79%	45.97%[a]	44.92%	40.66%	39.80%
	BiLSTM	55.51%[a]	55.85%	49.05%[a]	45.24%[a]	43.38%[a]	39.95%
SCC	-	**70.74%**	**65.61%**	**65.22%**	**57.83%**	**55.22%**	**47.52%**

- Explainability performarnce are reported in terms of precision, i.e. the degree of overlap between the top-N ranked sentences and the N-sentences annotators' summaries
- Annotators's summaries were built by grouping the sentences for which at least one, two or three out of the four annotators judged them among the most important ones
- The **SCC** model outperforms the **ExHiT** model
- Changing the **merging strategies** in the **ExHiT** architecture does not have a great impact on the performances

EXS4EXSA

CONCLUSIONS

DISCUSSION

- For the best of our knowledge, this is the **first attempt** to build a document classification model that generate an **extractive summary** in order to provide an easy to interpret explanation to the user.

- Both models have achieved good **Sentiment Analysis** results, not so far from the state of the art performance on the IMDB dataset.
  The classification task is accomplished while **performing an explanation** of the decision that may be considered particularly good for the SCC model.

- The attention weights rankings performed to build the extractive summaries for the **ExHiT** model has shown that it resulted particularly sensitive to the added empty sentences, that may be seen as an **additional noise**.

CONCLUSIONS
FUTURE WORKS

- The models were evaluated on a sentiment analysis task, but both architectures allow to use them in any **document classification task** (e.g. topic classification).

- Sentiment classification is a task that relies particularly on the sentences meaning in the document.
  This may be the main reason of **SCC** outperforming **ExHiT**.
  Evaluating the two models on a different kind of task may lead to a **fairer** comparison.

- Both models hide the potential to be able to operate on longer documents than regular transformers architectures.
  Thus, it would be interesting to evaluate their application on longer-documents tasks.

REFERENCES

[1] D. Gunning, Explainable artificial intelligence (xai), Defense Advanced Research Projects Agency (DARPA), nd Web (2017).

[2] W. S. El-Kassas, et al., Automatic text summarization: A comprehensive survey, Expert Systems with Applications (2020) 113679.

[3] A. Vaswani, et al., Attention is all you need, in: Advances in neural information processing systems, 2017.

[4] J. Devlin, et al., Bert: Pre-training of deep bidirectional transformers for language under-standing, arXiv preprint arXiv:1810.04805 (2018).

[5] E. Voita, et al., Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, arXiv preprint arXiv:1905.09418 (2019).

[6] L. Franz, et al., A deep learning pipeline for patient diagnosis prediction using electronic health records, arXiv preprint arXiv:2006.16926 (2020).

[7] R. Pappagari, et al., Hierarchical transformers for long document classification, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019.

[8] X. Zhang, et al., Hibert: Document level pre-training of hierarchical bidirectional transformers for document summarization, arXiv preprint arXiv:1905.06566 (2019).

[9] A. Maas, et al., Learning word vectors for sentiment analysis, in: Proceedings of the 49th annual meeting of the association for computational linguistics: Human languagetechnologies, 2011, pp. 142–150.

[10] K. Krippendorff, Estimating the reliability, systematic error and random error of interval data, Educational and Psychological Measurement (1970).

EXS4EXSA

THANKS FOR YOUR ATTENTION!

X-SENTIMENT

6TH INTERNATIONAL WORKSHOP ON
EXPLAINABLE SENTIMENT MINING AND EMOTION DETECTION

EUROPEAN SEMANTIC WEB CONFERENCE 2021 - ESWC 2021

EXS4EXSA

L. BACCO - A. CIMINO - F. DELL'ORLETTA - M. MERONE

JUNE 7, 2021