TERMS OF SERVICE SUMMARIZATION

E. RUFFOLI, F. HUDEMA, T. BALDI

CLICK-TO-AGREE

Problem

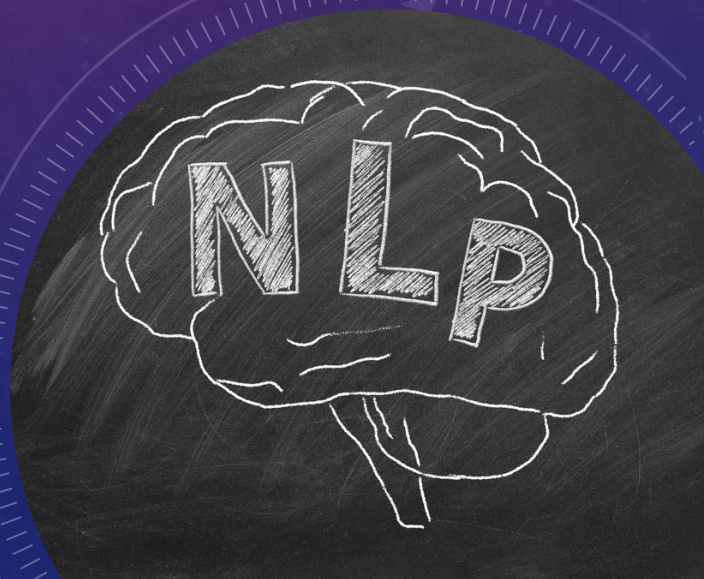
- Users give the right to keep, analyze and sell their data to web-based services and third parties by accepting services contracts about which they are not aware of. Increasingly often, **people click away their right to go to court if anything goes wrong**.

Solution

- Automated synthesis system that will extract key information from contracts in order to make them more understandable to the users.

How

- The service we offer is a software that uses algorithms to summarize textual documents. This specific field of Artificial Intelligence is called NLP and today is certainly one of the branches on which more research is carried out.



TEXT SUMMARIZATION

Extractive

Alice and Bob took the train to **visit the zoo**. They **saw** a baby giraffe, a lion and **a flock of** colorful tropical **birds.**

- **Summary**:

Alice and Bob visit the zoo saw a flock of birds

Abstractive

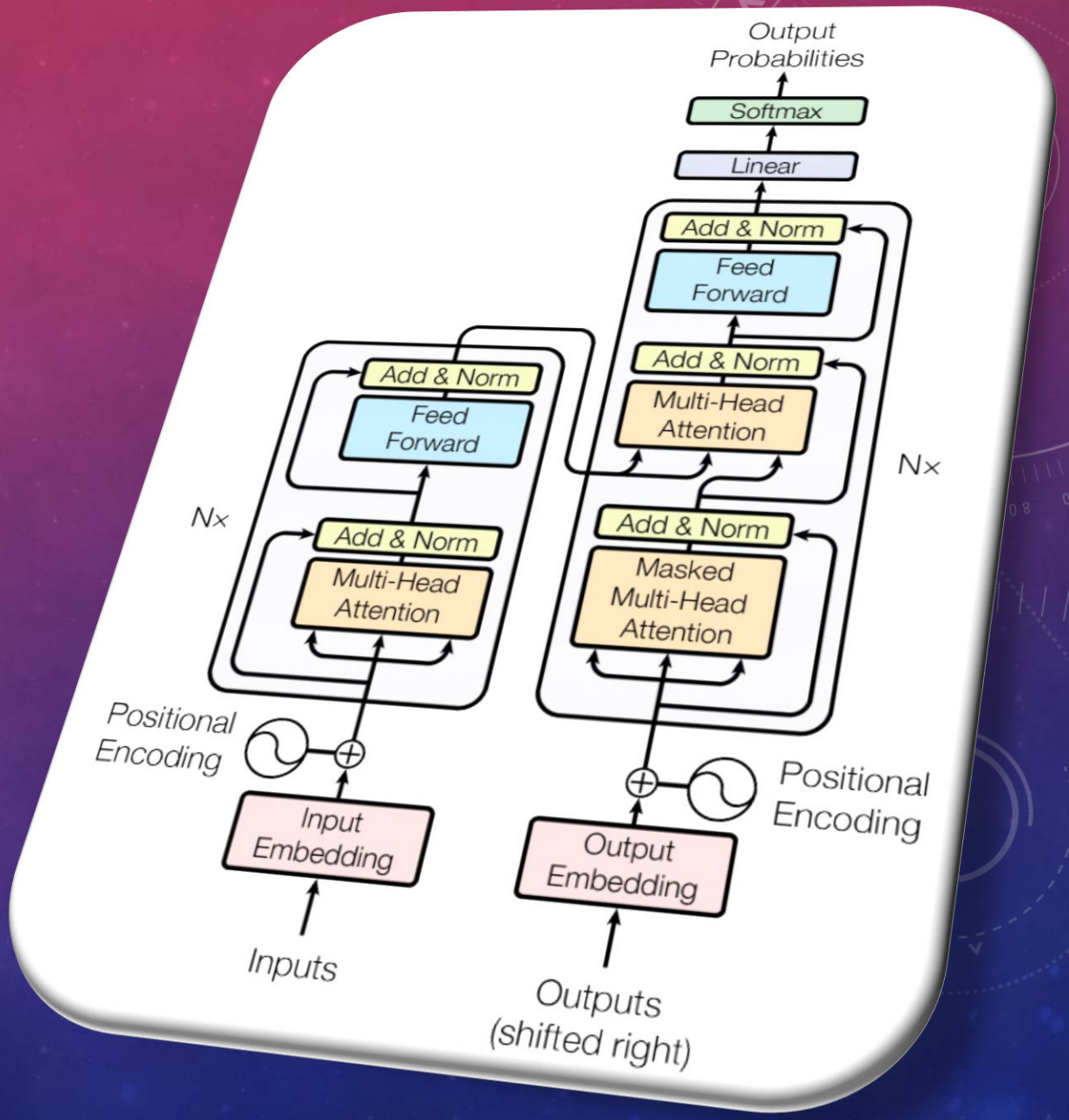
Alice and Bob took the train to visit the zoo. They saw a baby giraffe, a lion and a flock of colorful tropical birds

- **Summary**:

Alice and Bob **visited** the zoo **and** saw **animals** and birds

TRANSFORMERS

- Introduced by Google in December 2017 in the paper «Attention is all you need».
- Transformers turned out to be a gamechanger for natural language processing.
- They are based on the concept of «Attention»: a mechanism that prioritizes some parts of the input data. In this way the network can devote more focus to the important parts of the data.



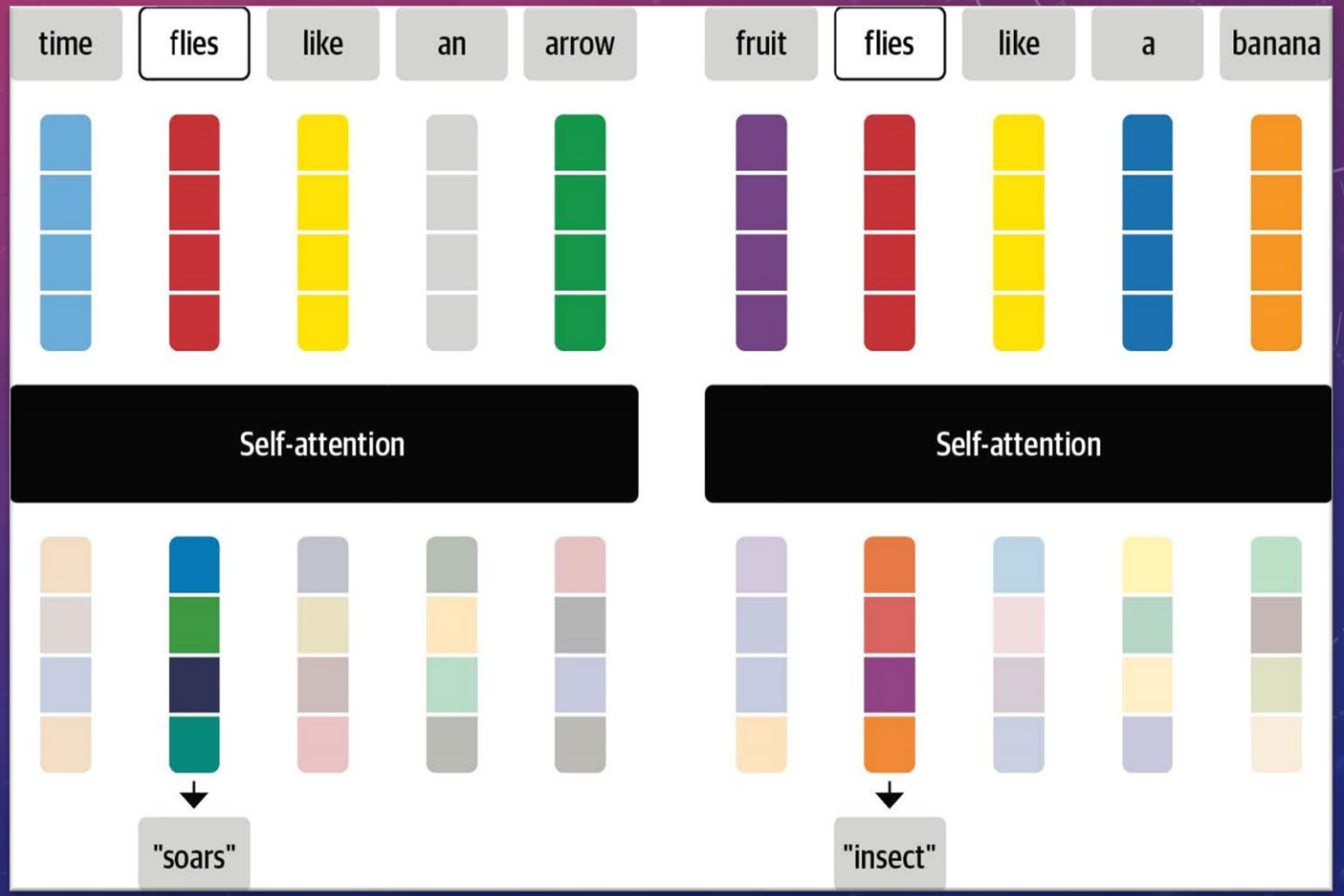
ATTENTION

For text sequences, the elements are token embeddings where each token is mapped to a vector of fixed dimension.

$$x'_i = \sum_{j=1}^{n} w_{ji} x_j$$

Attention updates raw token embeddings (x) into contextualized embeddings (x') to create representations that incorporate information from the whole sequence.

Attention can generate two different representations for the word "flies" based on the context.



SCRAPING

Collection of contracts from *tosdr.org*, through a Python script that exploits the API provided by the site.

Terms of services and privacy policies of major Internet services and companies.

Summaries of the main components of the contracts written by hand.

Uber Privacy Notice:

[…]We collect data when users create or update their Uber accounts. This may include their name, email, phone number, login name and password, address, profile picture, payment or banking information (including related payment verification information), driver's license and other government identification documents (which may indicate document numbers as well as birth date, gender, and photo). We may use the photos submitted by users to verify their identities, such as through facial recognition technologies. How we use personal data Location data (driver and delivery person): Uber collects this data when the Uber app is running in the foreground (app open and on-screen) or background (app open but not on-screen) of their mobile device. Location data (riders and delivery recipients). We collect precise or approximate location data from riders' and delivery recipients' mobile devices if they enable us to do so.[…]

Summary:

Many different types of personal data are collected. Your biometric data is collected. Information is provided about how your personal data is used. This service receives your precise location through GPS coordinates.

DATASET PREPROCESSING AND ANALYSIS

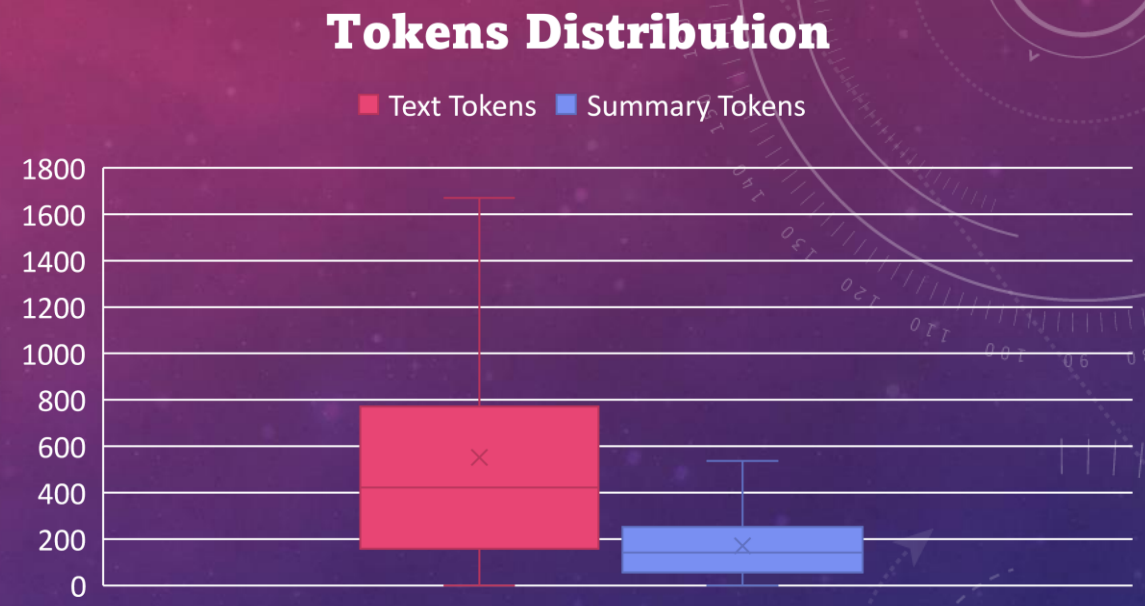
Raw dataset

Delete documents without summary

Remove any special character and the html tags

Discard non-English documents

901 terms of services and privacy policies contracts



Since summarization models have a **maximum context length** in terms of tokens, so we need to analyze the **size of the input and output** documents.

HOW TO FINE TUNE

Transformers model

- Transformer models (BART, T5, etc.) have been trained as language models. This means they have been trained on large amounts of raw text in a self-supervised fashion.

Fine-tuning

- Adaptation of the model to the summarization task. For our application we select pretrained models fine-tuned on English language that has been adapted to the summarization task using datasets as the *cnn_dailymail.*

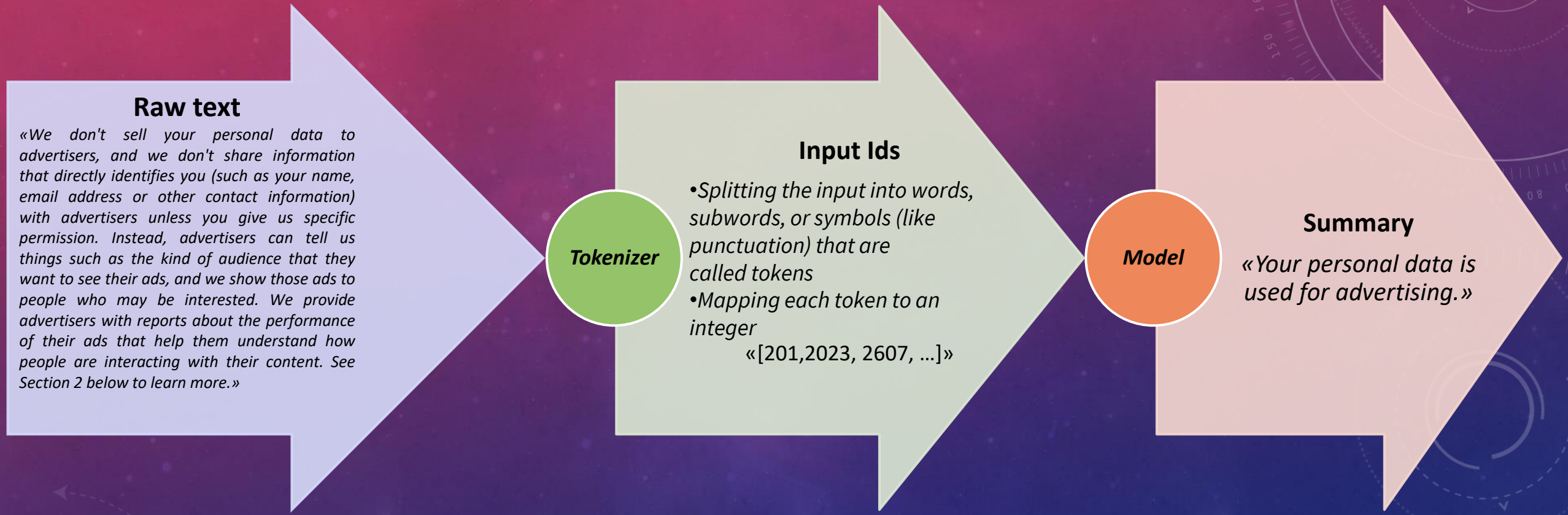
Training a Fine-tuned model

- Perform additional training with a dataset specific to the task. Fine-tuning in this step requires way less data and the amount of time and resources needed to get good results are much lower.

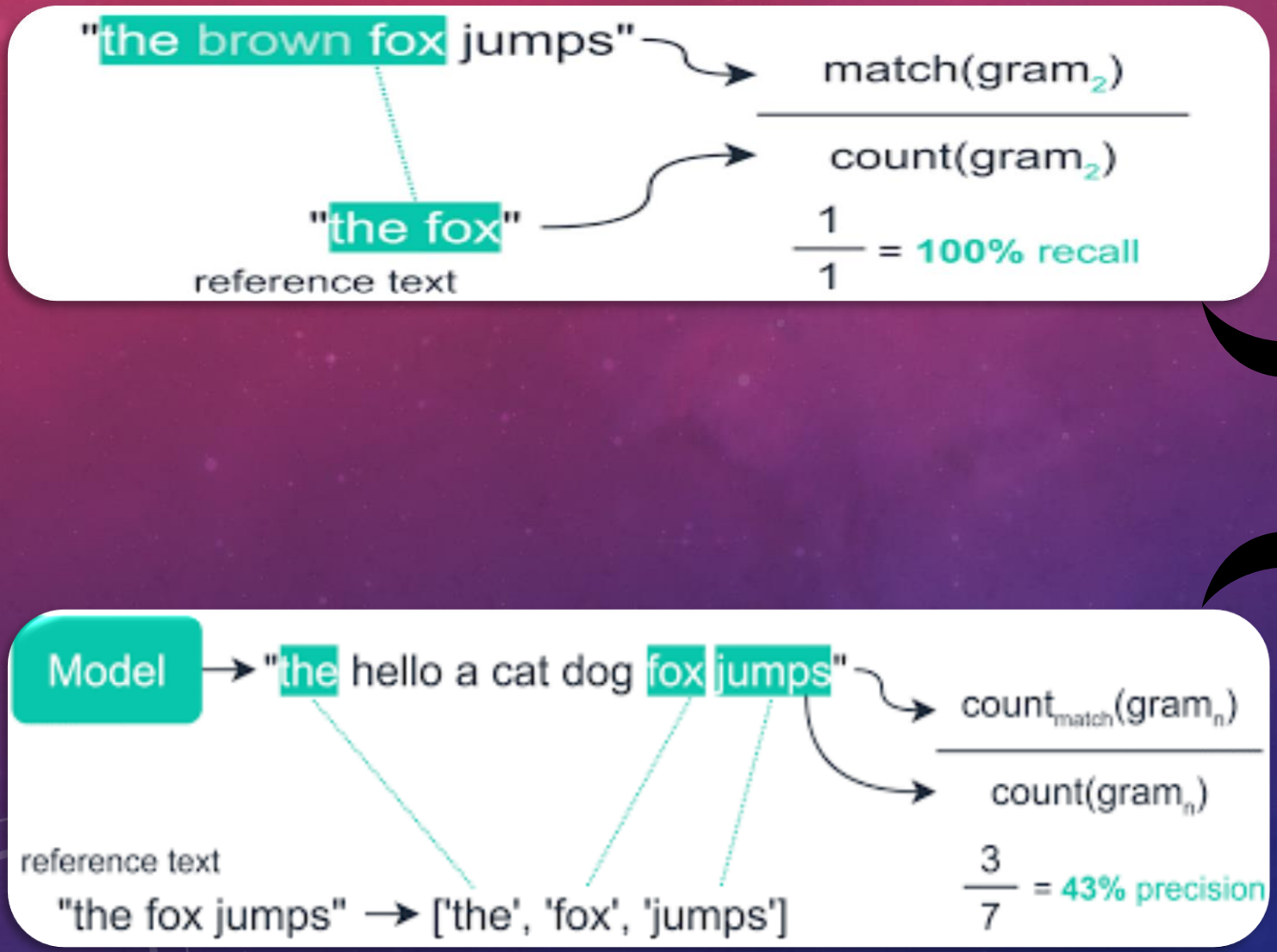
The Hugging Face ecosystem consists of two main parts: a family of **library** and the **Hub**. The library supports the major deep learning frameworks including PyTorch, and provides a standardized interface to a wide range of transofmer models.

Google Colab is a hosted **Jupyter notebook** service that requires no setup to use, while providing access free of charge to **computing resources** including GPUs.

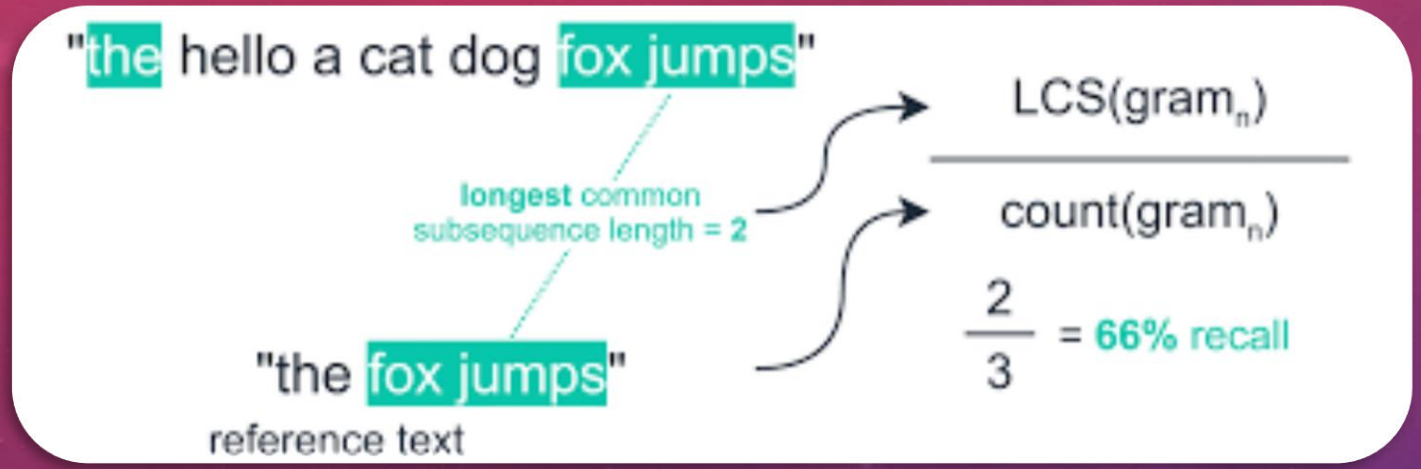
TRANSFORMERS PIPELINE



EVALUATION METRICS: ROUGE-N

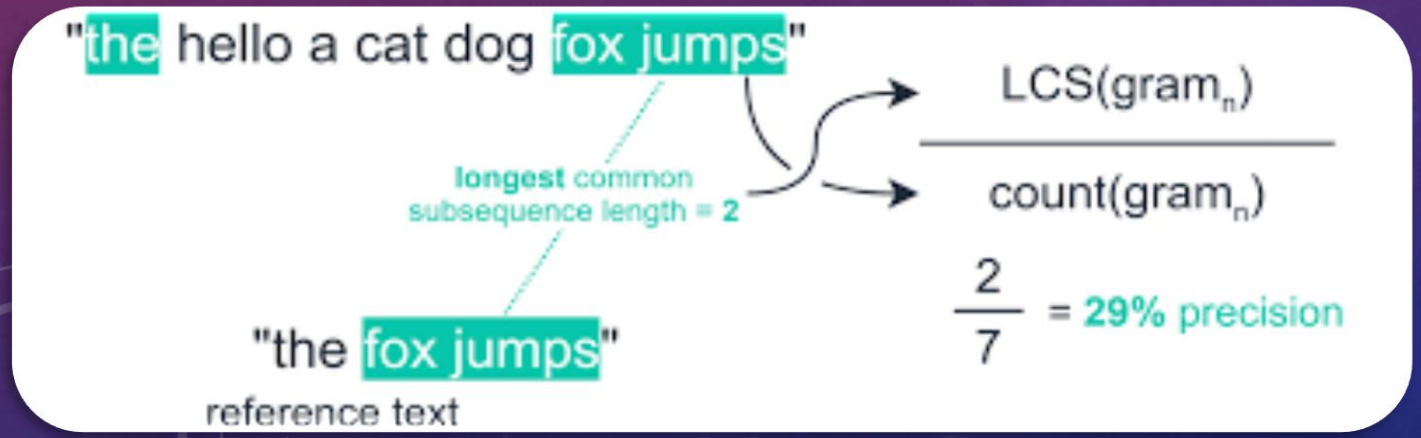


EVALUATION METRICS: ROUGE-L

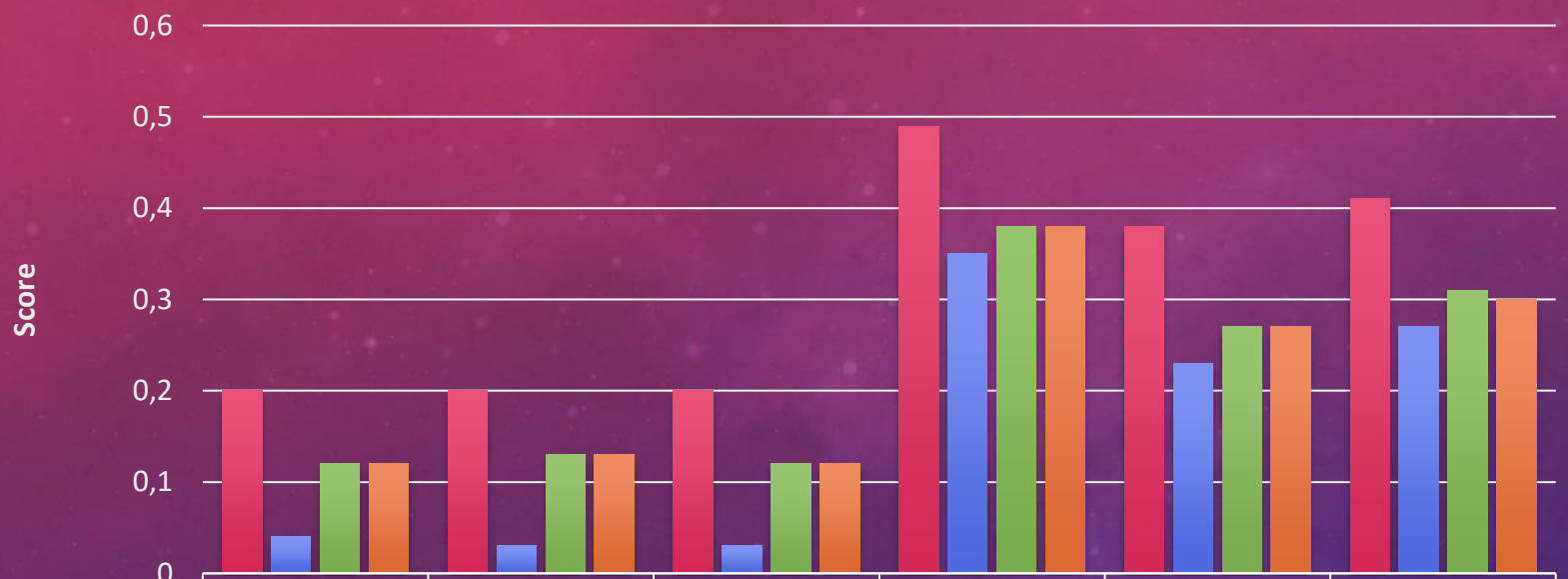


$$2 * \frac{0.29 * 0.66}{0.29 + 0.66} = 0.6$$

40% f1 score



MODELS COMPARISON

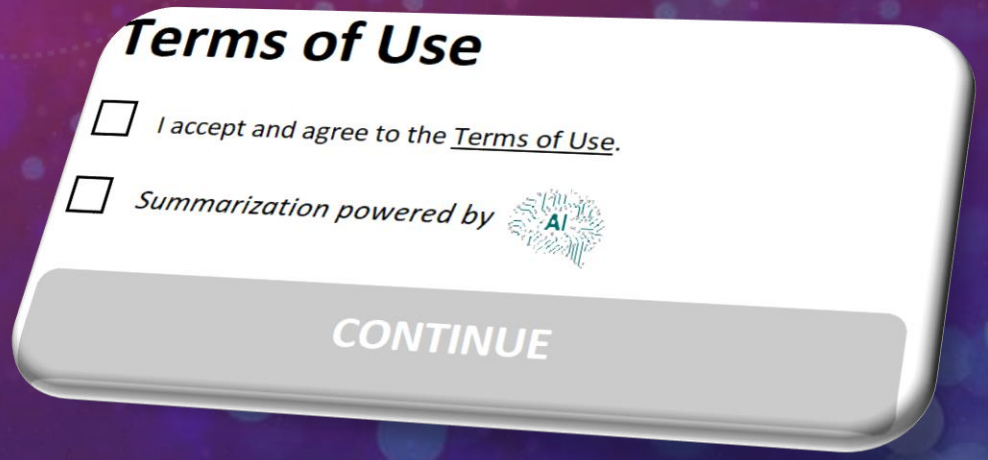


	Bart-cnn	T5-base	Pegasus-cnn	Bart-cnn-TOS	T5-base-TOS	Pegasus-cnn-TOS
rouge1	0,2	0,2	0,2	0,49	0,38	0,41
rouge2	0,04	0,03	0,03	0,35	0,23	0,27
rougeL	0,12	0,13	0,12	0,38	0,27	0,31
rougeLsum	0,12	0,13	0,12	0,38	0,27	0,3

- The limitation of these summarization models is that they have a maximum **input length** in terms of context tokens.
- Unfortunately, there is no single strategy to solve this problem, and it's an open and **active research question**.
- We adopted the solution to **split the text** in blocks smaller than the maximum number of tokens accepted as input by the models, being careful not to cut off sentences, and apply recursively the computation of summaries.

CONCLUSION

- Provide the user with a mechanism to summarize and extract key information from their contracts to be clear and transparent towards customers.



- The company can easily embed it in its web pages, in particular in contract boxes, in this way the user is forced to read the summarization in key points of the contract before continuing.

- Companies that operate in the digital industry and are configured as a hub of third-party services, in which users must accept new terms of service each time they purchase a service.

TRY IT!

Terms Of Service Summarizer Demo