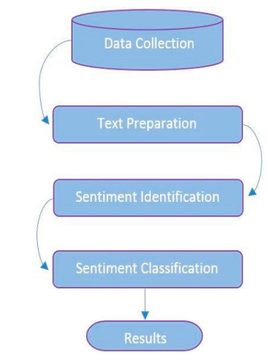
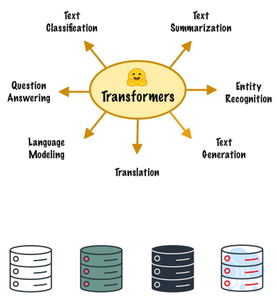
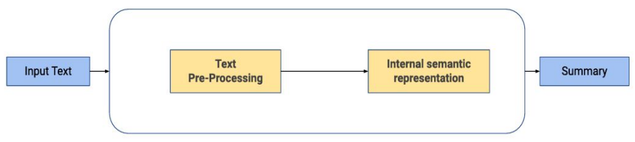
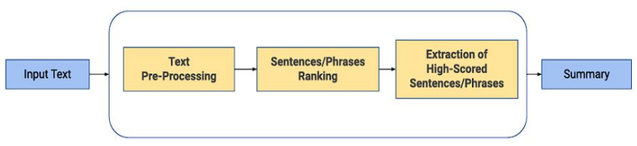
Artificial intelligence deep learning text summarization and sentiment analysis tool

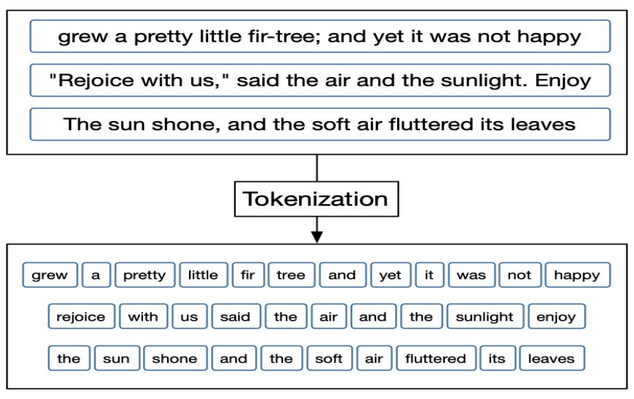
Danylo Fedorov - Emin Erkan Korkmaz
Department of Computer Engineering, Yeditepe University





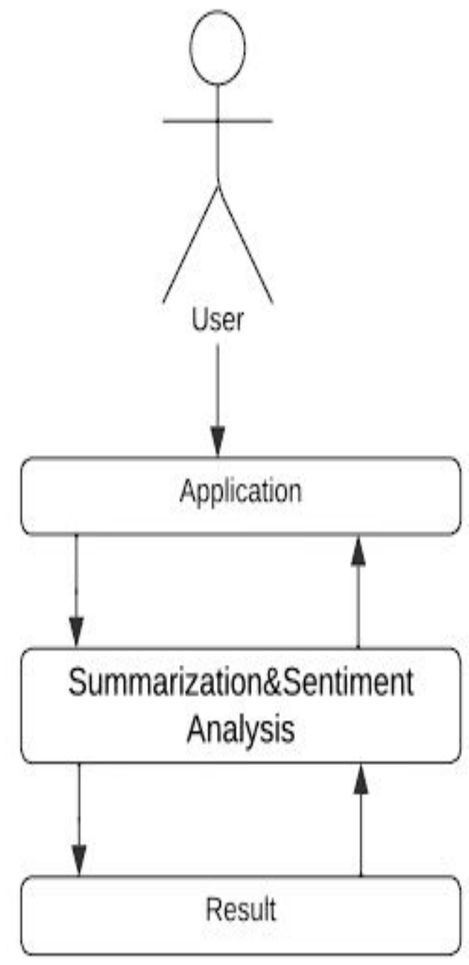






Abstract

Deep learning artificial intelligence text summarizer is the technique of reducing a given text to a shorter version, which can be performed in two ways: an extractive or an abstractive summarization.Extractive summarization, which can extract vital phrases and terms from the given text. The underlying the concept is to create a summary with the aid of selecting the maximum number of essential phrases from the given text.But on the other hand, there is Abstractive Summarization which involves the generation of totally new terms that seize the meaning of the input sentence. The underlying concept is to position a strong emphasis on the shape aiming to generate a grammatical precis thereby requiring advanced language modeling techniques. Both ways are used by humans in daily life.In addition to the use of the summarization technique, the sentiment analysis will be used and will be performed in two ways: by determining the polarity(Irrelevant, Negative,Positive, Neutral) and relevance(World, Sport, Business, Sci/Tech) of the text. By analyzing the sentiment of the text, we can achieve a better understanding of the emotions and opinions expressed within the given content.Some problems aspects of human life can be solved by deep learning artificial intelligence, which can make life a little bit easier. This paper, will be focused on the better way for text summarization via deep learning artificial intelligence, a deep dive into this topic and all aspects that are required to implement it. In conclusion, will be explained the methodologies that were used and what was achieved, and how it can be improved.



CNN Model

The chosen model architectures for text classification is CNN (Convolutional Neural Network) model. CNN architecture is particularly suitable for sentiment analysis, as it can learn to recognize patterns in text that indicate positive or negative sentiment by analyzing the local relationships between words or phrases. The CNN model for text classification consists of convolutional layers, pooling layers, and fully connected layers. CNN model ability to capture local patterns and hierarchical representations in text makes it suitable for understanding sentiment cues in natural language.

NLP Model

The chosen model architectures for text summarization is NLP (Natural Language Processing) model. NLP text summarization architectures, especially those based on Transformer models, have shown excellent performance and are considered the state-of-the-art for abstractive and extractive summarization.To successfully train the models, it is crucial to preprocess the input data, transforming it from a human-readable format to a machine-readable format. This preprocessing step helps the model understand the textual information and facilitates the training process.

TRAINING OF THE MODELS

For training NLP and CNN models public datasets of Kaggle and Huggingface sites are used. Before using the dataset, data inside was preprocessed from human-readable to machine-readable format. Datasets for each model were split into training and test sets with 10% data for testing and 90% data for training. As a result, each model was trained with a big amount of high-quality data.CNN model was trained with 200.000 examples and on the other hand NLP model was trained with 3 custom datasets with over 530.000 examples. Which makes Models well-suited for Sentiment Analysis and Text Summarization Process.The number of times the Text summarization and Sentiment Analysis learning algorithm sees each sample in the training set is around 30 epochs.

Results

Relevance Results

CLASS	TIME	AVG%
WORLD	0.0317	99.9%
SPORT	0.0324	100%
BUSINESS	0.0394	99.8%
TECH	0.3234	99.8%
AVG:	0.0339	99.91%

Polarity Results

CLASS	TIME	AVG%
IRRELEVANT	0.0325	81.0%
NEGATIVE	0.0324	95.4%
POSITIVE	0.0352	69.74%
NEUTRAL	0.0328	93.6%
AVG:	0.0337	84.97%

Abstractive model accuracy

Sumary	Precision	Recall	F1
1	91	90	90
2	93	92	90
3	92	92	92
4	86	81	83
5	93	92	93
AVG:	91	89.4	90

Extractive model accuracy

Sumary	Precision	Recall	F1
1	91	89	90
2	81	80	81
3	83	81	82
4	80	76	78
5	77	83	84
AVG:	82.4	81.8	83

Extractive model performance

Test	MinTime	MinLen	OriginalWorldCount	MaxTime	MaxLen
1	4.6	28	450	10.8	88
2	6.6	35	698	12.1	104
3	6.0	33	738	12.5	100
4	5.2	34	414	10.7	97
5	5.9	30	705	12.3	94
AVG:	5.6	32	601	11.68	96.6

Abstractive model performance

Test	MinTime	MinLen	OriginalWorldCount	MaxTime	MaxLen
1	7.4	39	364	8.8	57
2	7.3	25	335	9.7	43
3	7.0	40	526	7.8	58
4	8.4	46	404	9.7	62
5	8.1	35	453	7.4	55
AVG:	7.6	37	394.4	8.9	55

Future Work

The goal is to enhance the fine-tuned T5-base model by training it with diverse language datasets. This will create a powerful summarization tool that allows users to choose any language and desired summary length. To achieve this, it is recommended to train the model with a larger dataset that includes a wide range of patterns, enabling it to capture complex patterns and dependencies in the input text. This will result in improved performance across various natural language processing tasks.

References

1. Quillbot. “Quillbot - ai paraphrasing tool.” (2023).Available: bit.ly/3Xcamit.
2. MonkeyLearn. “Monkeylearn - sentiment analysis online.” (2023). Available: https://monkeylearn.com/sentiment-analysis-online/.
3. M. Lewis, Y. Liu, N. Goyal, et al., “BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,”CoRR, vol. 1910.13461, pp. 7871–7880, 2019.

Contact

erkan.korkmaz@yeditepe.edu.tr
danylo.fedorov@std.yeditepe.edu.tr