Data Augmentation using Transformers and Similarity Measures.

Mustafa Adel Ibrahim · Following
6 min read · May 12

Arabic Textual Data Augmentation Using AraGPT-2, AraBERT & Similarity Measures to increase dataset size, variability and solve imbalanced class.

Abstraction.

Models can achieve high efficiency & better performance upon training on **big datasets.** But, obtaining a large amount of labeled data is difficult, especially when developing AI applications in domains such as *education, healthcare, etc.*

In this view, massive research exists in the literature to address the dataset adequacy issue. One promising approach for solving dataset adequacy issues is *data augmentation (DA).*

What's DA? Is it fruitful?

- DA is to intelligently increase the dataset size by making different transformations on the available instances to generate new correct & representative ones.

- DA increases the dataset size, its variability in addition to solving the class imbalance problem in order to well-generalize. More over, it's considered a way to minimize overfitting.

> ***DA is well-established in computer vision. But "why it's not a common practice in NLP"?!***

- Unfortunately***, not all augmentation methods are applicable to Arabic.*** *Due* to *the Arabic language's unique characteristics*. the applicable transformation to a specific language is not necessarily to be applied to the Arabic textual data.
- Increasing Arabic textual data is based on traditional ways. But using traditional ways is ***both costly & time-consuming***, especially when there're not enough resources to support the augmentation process.
  For instance, lacking enough language dictionaries/thesaurus /vocabulary or a database of synonyms for a given dataset.

> ***What're traditional DA methods?***

Previously, we've posted about traditional techniques, such as easy data augmentation & back translation. But let's recap briefly.

- Synonym replacement **(paraphrasing-based)**: In this technique, we replace some words in the sentence with their synonyms to help model handle variations in a language.
- Random insertion/deletion **(adding some noise)** : In this technique, we randomly insert/ delete/ swape/ subsstitute words in the sentence. This can help the model learn to handle noisy data and improve its robustness.

Motivations Behind this Paper.

> *- Arabic's considered the world's 5th-most-widely spoken language and there's a high growth of Arabic content on the Internet.*
> *- The results show that DA's effectiveness in the English language's learning techniques. But there's a lack of research on DA for Arabic.*
> *- Nevertheless, none of traditional techniques correctly augment the Arabic textual data.*

Transformers.

Transformers are a type of neural network architecture that has revolutionized NLP. They can capture contextualized information & relationships between words in a sentence, making them effective in different NLP tasks, including text summarization, translation & question-answering systems in addition to challenges in different languages. Besides, employing them while dealing with text reduces the training time & computations.

As a result, transformer models can be used for DA as a paraphrasing-based technique while also preserving the context of the augmented data instances.

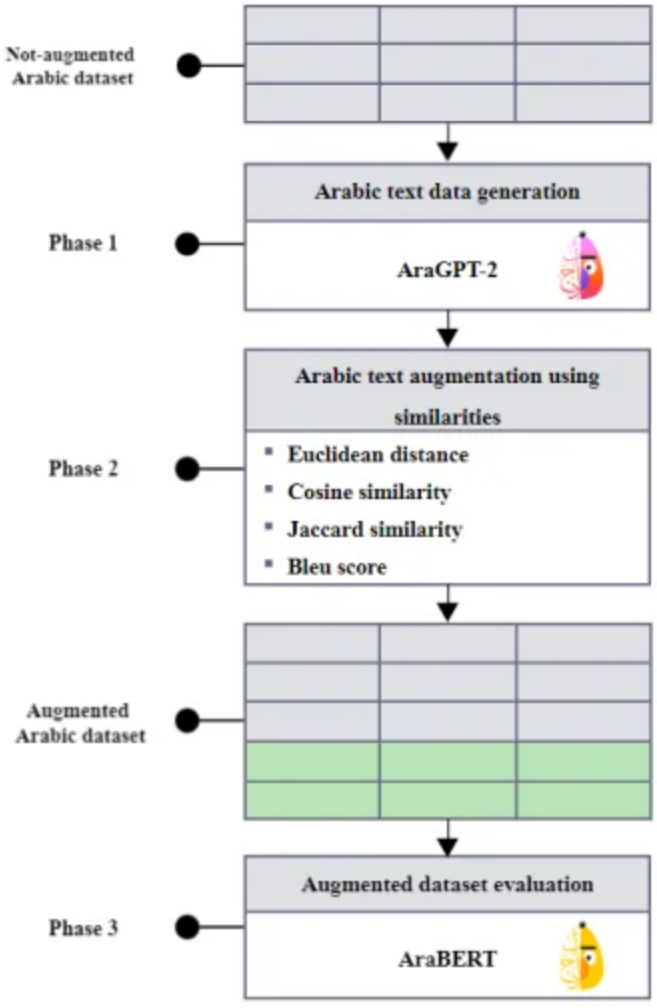
Methodology.

This Paper proposed a "3-phase empirical approach".

- **In the 1st phase**, the AraGPT-2-base pre-trained model is used to generate Arabic text from the given-dataset records.
- **In the 2nd phase**, new records are to be added to the dataset by employing the **similarity measures**.

This process is dependent on 2 measures which are the **similarity-thresholds** & **the selection of *class-label*** *[imbalanced class]*.

- **The 3rd phase** (a complementary phase) it assists in evaluating the performance of the text classification process on the newly created dataset (i.e., the given + the augmented dataset).
- **Finally**, we used the AraBERT transformer to do the classification tasks for evaluating the proposed approach on the augmented Arabic textual dataset's effect on performance.

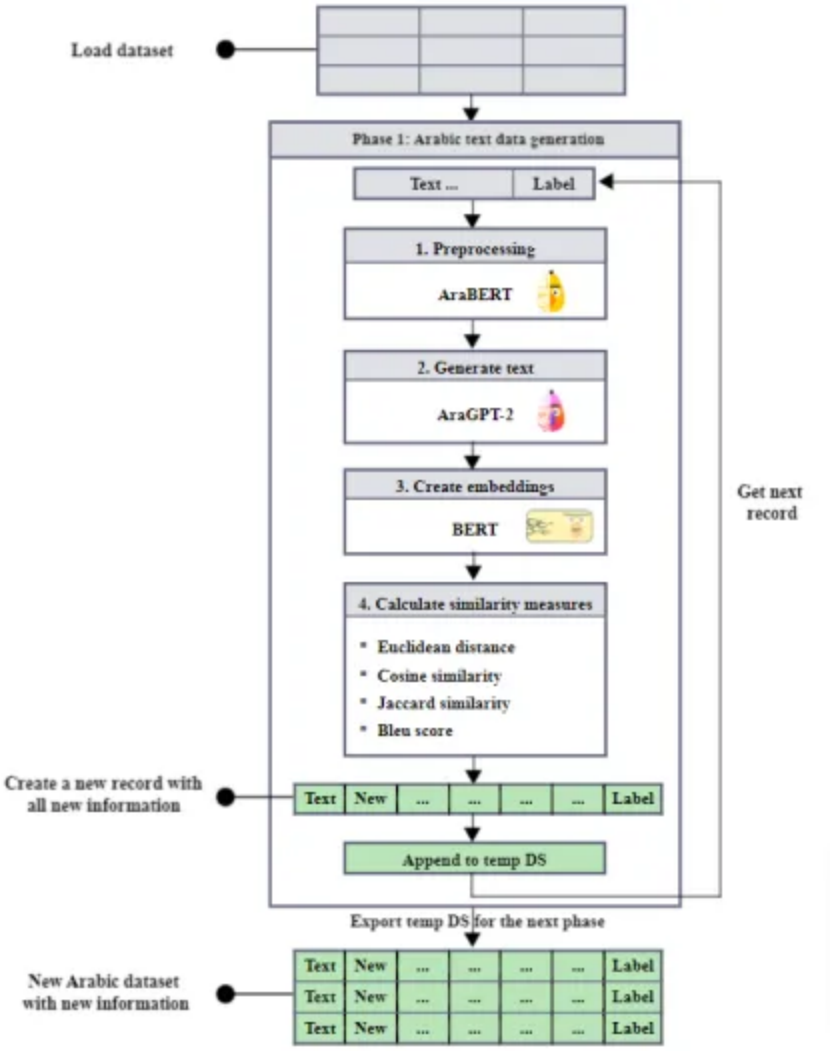


The main 3-phases of the adopted methodology

> *Although employing* ***transformer models preserve the text context****, it's essential for DA to evaluate the sentence before adding it to the data to ensure that this augmentation will improve the performance without harming the original data.*
> *So, the quality of augmented text should be assessed from various perspectives in terms of* ***context, semantics, diversity, & novelty****.*
> *In this sense,* ***text-similarity metrics (i.e., Euclidean, cosine, Jaccard, BLEU score)*** *can be used to check their quality.*

PHASE-1: ARABIC TEXT DATA GENERATION USING TRANSFORMERS

First-thing-first, the dataset to be augmented is loaded. Then, create a transformer that can generate Arabic text (AraGPT-2) along with initializing the similarity function/s needed to calculate the similarity between the old supplied and the newly generated text.



Methodology-steps contained within phase 1

Then, for each record in the given dataset's records:

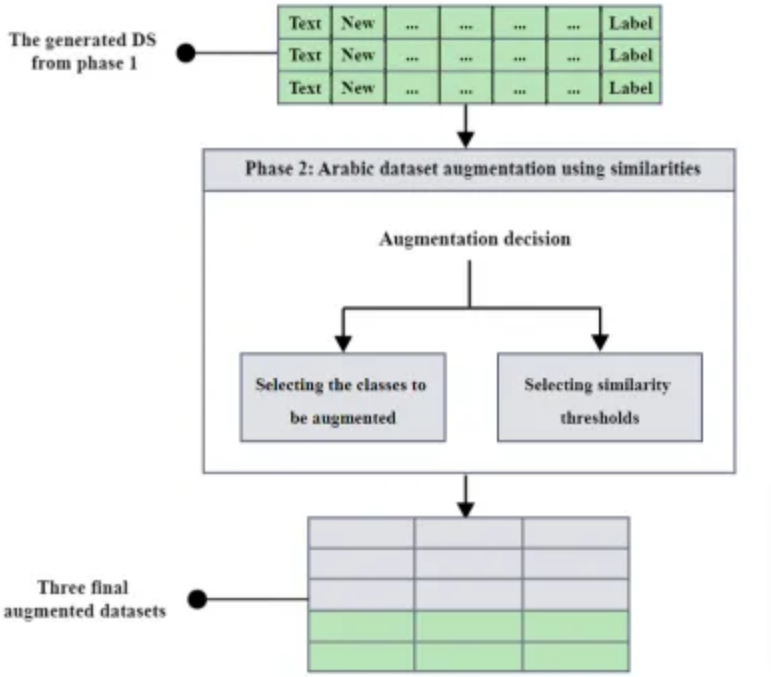
- **Firstly**, the given text in is pre-processed (AraBERT-pre-processor).
- **Secondly**, calculate the word embedding (WE) that represents the given Arabic text would take place. Such a sub-step is needed since the similarity functions deal with numerical representations (vectors) rather than the abstract Arabic text representation to calculate the distances between the objects for comparing them.
  "Hence, the authors used BERT WE for computing WE."

- **Thirdly**, the similarity between the numerical representation of the given Arabic text and the newly generated one is calculated with the selected similarity functions (i.e., Euclidean, cosine, Jaccard, and BLEU distances).
- **Finally**, all the computed and generated information were collected within the current loop (the given Arabic text, the related class label, the newly generated text, all text of the given Arabic text combined with the generated one, the embedding representation, and the similarities' values), and is appended to the dataset.

Moreover, such a record is added to the final dataset to be exported upon finishing this phase, along with the original class label related to the current record being processed.

PHASE-2: ARABIC DATASET AUGMENTATION USING SIMILARITIES

the generated dataset from the previous phase is processed to generate one final dataset that contains the new augmented records. Achieving this final dataset requires 2 significant decisions:



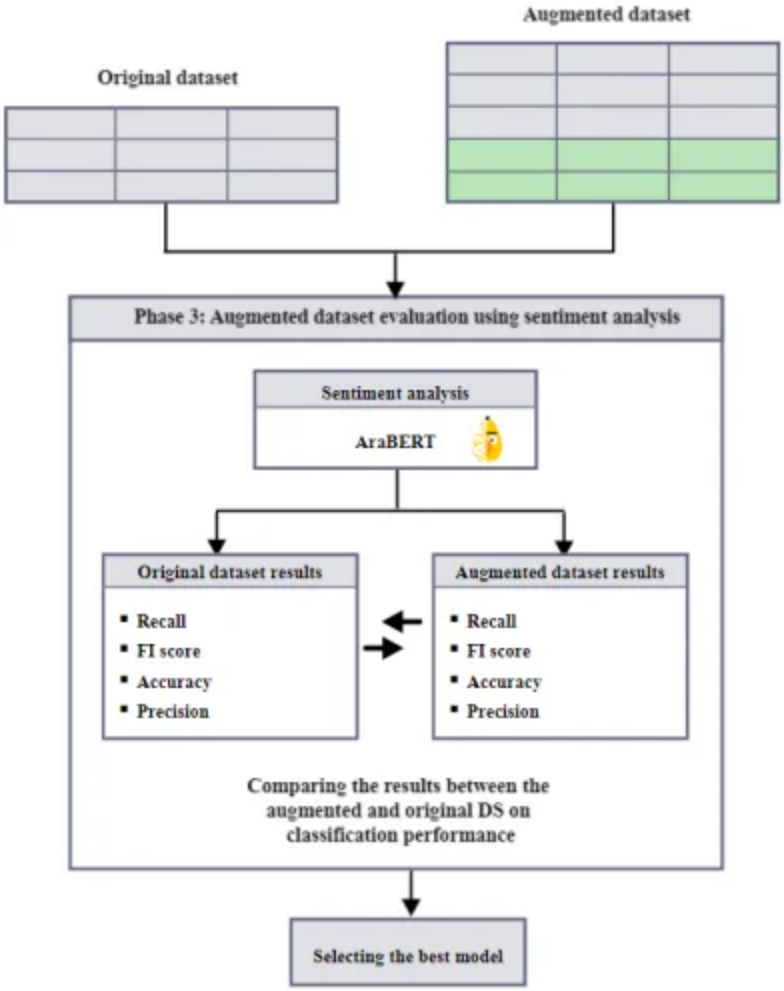
Methodology-steps contained within phase 2.

- The *1st decision* is selecting the classes to be augmented.
- The *2nd decision* is selecting a threshold (i.e., similarity-desired value) to decide the selection process of the newly generated text as a new record in the new dataset along with the related class label.

PHASE-3: Augmented dataset evaluation using sentiment.

The final dataset (i.e., the original + the augmented dataset) is evaluated using sentiment analysis (i.e., all selected datasets are classified with the sentiment of the text).
To conclude this evaluation:

- The model of the original dataset (i.e., before augmentation) is needed to find the classification performance (i.e., recall, accuracy, etc.).
- Compare the obtained results with the results found in the same classification process on the augmented dataset (i.e., the original + augmented dataset).



Methodology-steps contained within phase 3

Datasets

- Ara_Sarcasm is for detecting sarcasm in Arabic.
- ASTD is collected from tweets after being filtered and annotated by the authors to be an Arabic social sentiment analysis dataset.
- ATT is for the reviews expresses the attraction sentiment of the travelers. collected from TripAdvisor.
- MOVIE is also scrapped TripAdvisor to rate watched movies.

Dataset Name	Record No.	Class Labels' Information		
		Label Name	No. Instances	Ratio (%)
Ara-Sarcasm	10545	POSITIVE	1678	15.91 %
		NEUTRAL	5339	50.63 %
		NEGATIVE	3528	33.46 %
ASTD	3221	POS	776	24.09 %
		NEUTRAL	805	24.99 %
		NEG	1640	50.92 %
ATT	2151	POS	81	3.77 %
		NEG	2070	96.23 %
MOVIE	1517	POS	966	63.68 %
		NEUTRAL	170	11.21 %
		NEG	381	25.12 %

The considered dataset for experimentations in the proposed solutions.

Results.

Dataset augmentation and growth percentage.

The growth percentage is measured concerning the total number of the original instances in that set.

GROWTH COUNTS – ATT DATASET

Class Labels	Dataset Type				
	Original Dataset	cosine Dataset	Euclidean Dataset	Jaccard Dataset	BLEU Dataset
POS	81	86	116	128	126
NEG	2070	2070	2070	2070	2070
Total	2151	2156	2186	2198	2196

GROWTH COUNTS – ARASARCASM DATASET

Class Labels	Dataset Type				
	Original Dataset	cosine Dataset	Euclidean Dataset	Jaccard Dataset	BLEU Dataset
NEGATIVE	3528	5245	4846	5317	5275
NEUTRAL	5339	5339	5339	5339	5339
POSITIVE	1678	2607	2262	2459	2672
Total	10545	13191	12465	13115	13286

GROWTH COUNTS - MOVIE DATASET

Class Labels	Dataset Type				
	Original Dataset	cosine Dataset	Euclidean Dataset	Jaccard Dataset	BLEU Dataset
POS	381	381	556	762	617
NEG	170	170	247	340	291
NEUTRAL	966	966	966	966	966
Total	1517	1517	1769	2068	1874

GROWTH COUNTS - ASTD DATASET

Class Labels	Dataset Type				
	Original Dataset	cosine Dataset	Euclidean Dataset	Jaccard Dataset	BLEU Dataset
NEG	1640	1640	1640	1640	1640
NEUTRAL	805	1074	1209	1134	1238
POS	776	1072	1134	1110	1237
Total	3221	3786	3983	3884	4151

the results of this experiment are summarized for each dataset.

Resources

Paper: [2212.13939] Data Augmentation using Transformers and Similarity Measures for Improving Arabic Text Classification (arxiv.org)

Pervious post: Textual data augmentation

NLP | Data Augmentation | Transformers

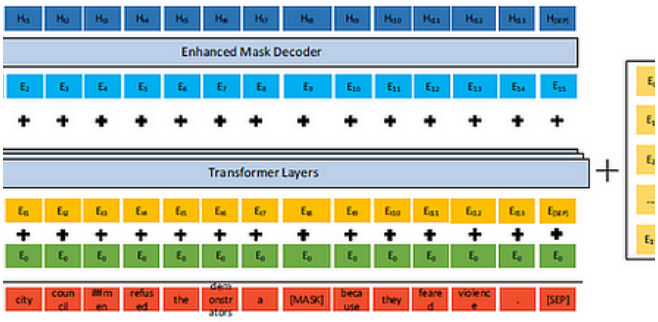
Written by Mustafa Adel Ibrahim

Following

11 Followers

Data Scientist | AI Engineer. https://www.linkedin.com/in/mustafa604/

More from Mustafa Adel Ibrahim

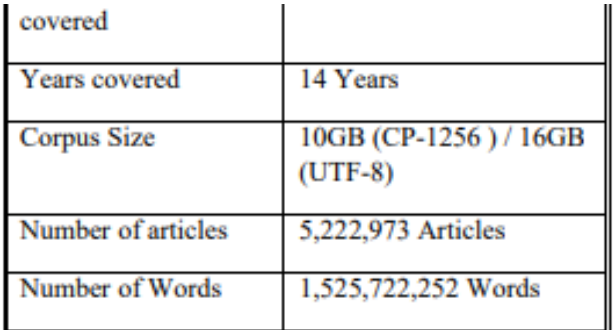


Mustafa Adel Ibrahim

DeBERTa: Decoding-enhanced BERT with Disentangled Attention

DeBERTa, 1st Single Model to Surpass Human Performance on SuperGLUE

6 min read · May 29



covered	
Years covered	14 Years
Corpus Size	10GB (CP-1256) / 16GB (UTF-8)
Number of articles	5,222,973 Articles
Number of Words	1,525,722,252 Words

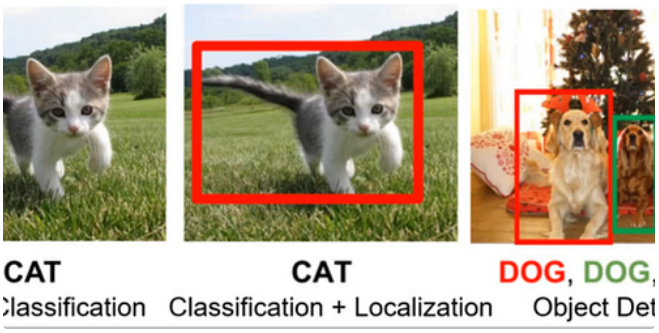
Mustafa Adel Ibrahim

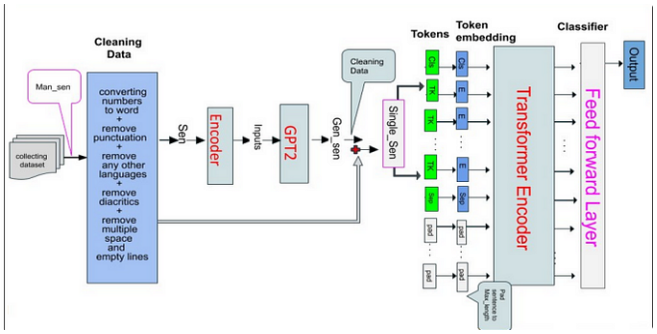
1.5 billion words Arabic Corpus

Building a contemporary linguistic corpus for Arabic language from different countries...

5 min read · Sep 16

6





Mustafa Adel Ibrahim

Deep Learning Specialization C4W3- Part1

Object detection is one of areas that's exploding and working so much better than...

17 min read · Aug 25

Mustafa Adel Ibrahim

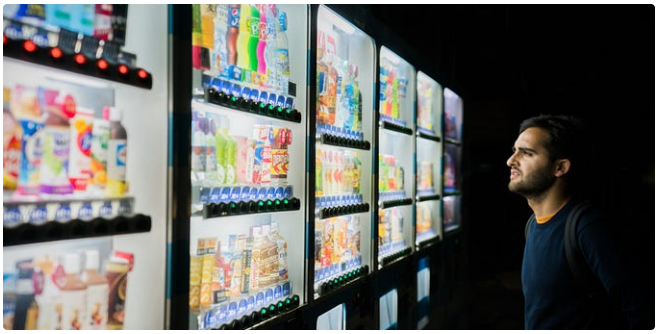
Detect Deepfake Text.

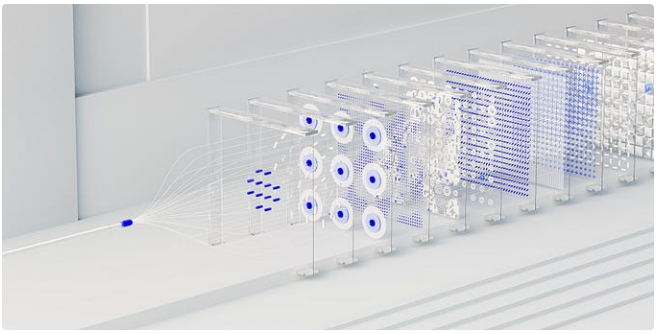
Automatically detect generated deepfake text using GPT2-Small-Arabic to distinguish...

4 min read · May 28

See all from Mustafa Adel Ibrahim

Recommended from Medium





Ovbude Ehi

Recommendation Engines

Practical Techniques: Content-Based Filtering, Collaborative Filtering and...

11 min read · May 8

5

Lefteris Charteros

Building a Sentiment Analysis Classifier using PyTorch Lightning

The purpose of this tutorial is to build a complete machine learning system that will...

17 min read · Jul 22

Lists

Natural Language Processing

669 stories · 283 saves

The New Chatbots: ChatGPT, Bard, and Beyond

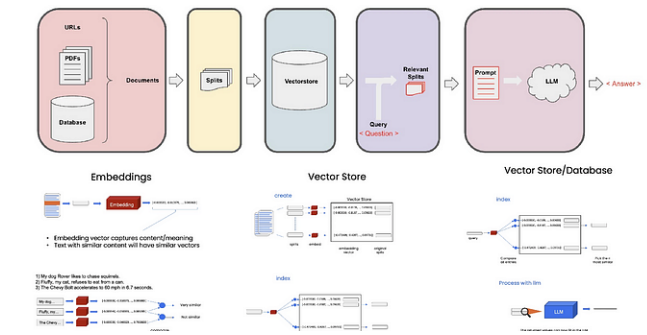
13 stories · 133 saves

New_Reading_List

174 stories · 133 saves

Staff Picks

465 stories · 317 saves

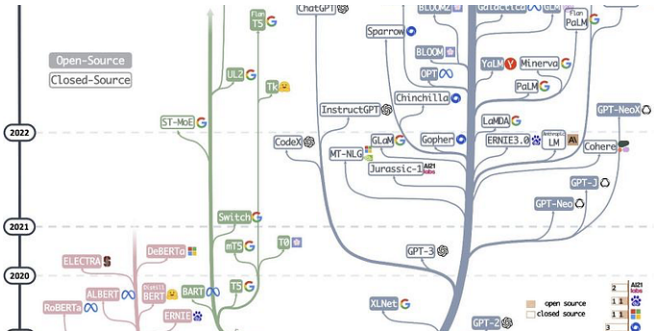


TeeTracker

Chat with your PDF (Streamlit Demo)

Conversation with specific files

4 min read · Sep 15



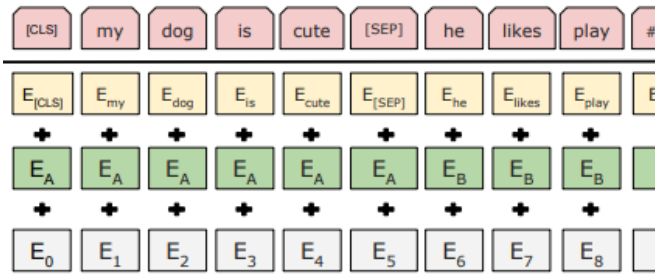
Haifeng Li

A Tutorial on LLM

Generative artificial intelligence (GenAI), especially ChatGPT, captures everyone's...

15 min read · Sep 14

56 372



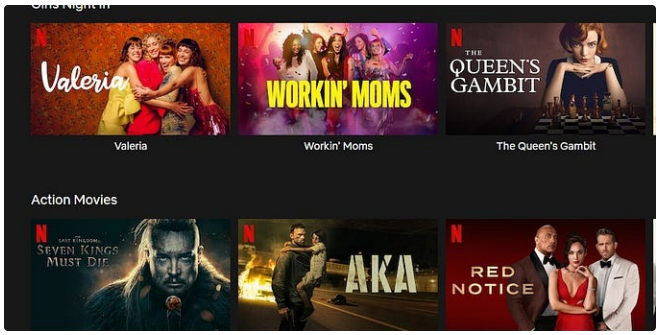
Zain ul Abideen

A Comparative Analysis of LLMs like BERT, BART, and T5

Exploring Language Models

6 min read · Jun 26

20 1



Deepa Pandit

Recommender System : User Collaborative filtering

by Deepa Pandit

10 min read · Jul 8

16

See more recommendations