ARABIC AUTOCOMPLETE SYSTEM

AGENDA

Preprocessing

Pre_Train Model

User Interface(GUI)

Evaluation

Search

JM100 AND 1 COLLABORATOR · UPDATED 4 YEARS AGO

10 Code Download

MNAD : Moroccan News Articles Dataset

Moroccan News Articles Dataset

Data Card Code (2) Discussion (0) Suggestions (0)

About Dataset

The MNAD corpus is a collection of over **1 million Moroccan news articles** written in modern Arabic language. These news articles have been gathered from 11 prominent electronic news sources. The dataset is made available to the academic community for research purposes, such as data mining (clustering, classification, etc.), information retrieval (ranking, search, etc.), and other non-commercial activities.

Dataset Fields

Usability
8.24

License
CC0: Public Domain

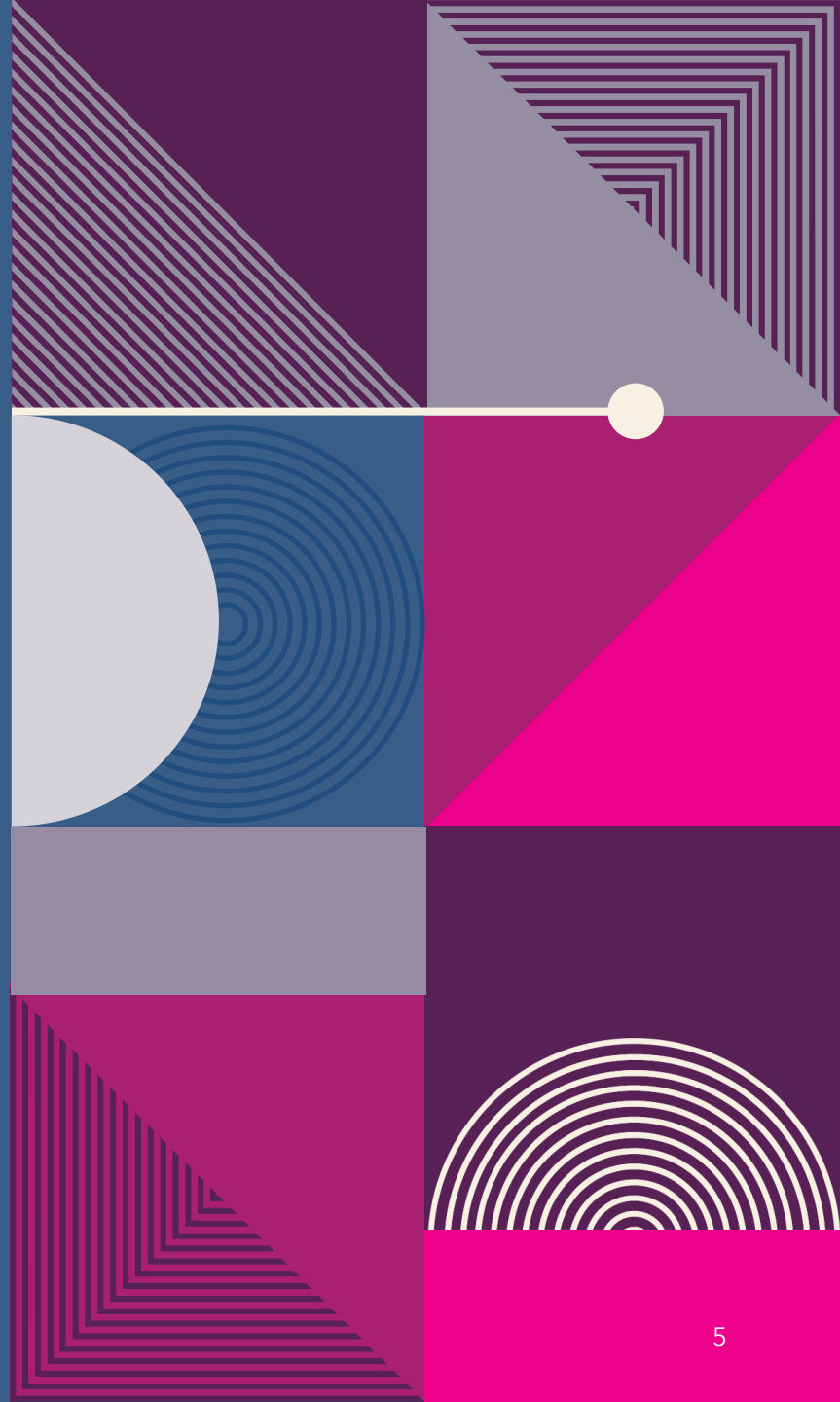
Expected update frequency
Annually

Tags

PREPROCESSING

TEXT DATA PREPARATION FOR ANALYSIS

- It removes web addresses (URLs) that might be present in the text.
- It eliminates HTML tags, which are often found in text extracted from web pages.
- It strips out any characters that are not standard alphanumeric characters or Arabic characters, ensuring that the text contains only relevant information.
- It removes all digits from the text.



Model Used
aubmindlab/aragptesab-2
A GPT- depoleved ledom egaugnal cibarA desab-2
.baL dniM barA eht yb

Why This Model?

☑ **Arabic Language Specialized**

Trained on a large-scale Arabic corpus, making it ideal for understanding and generating high-quality Arabic text.

☑ **Causal Language Modeling**

Designed for *auto-regressive* tasks —it predicts the next word based on the previous context, perfect for text generation.

☑ **Hugging Face Integration**

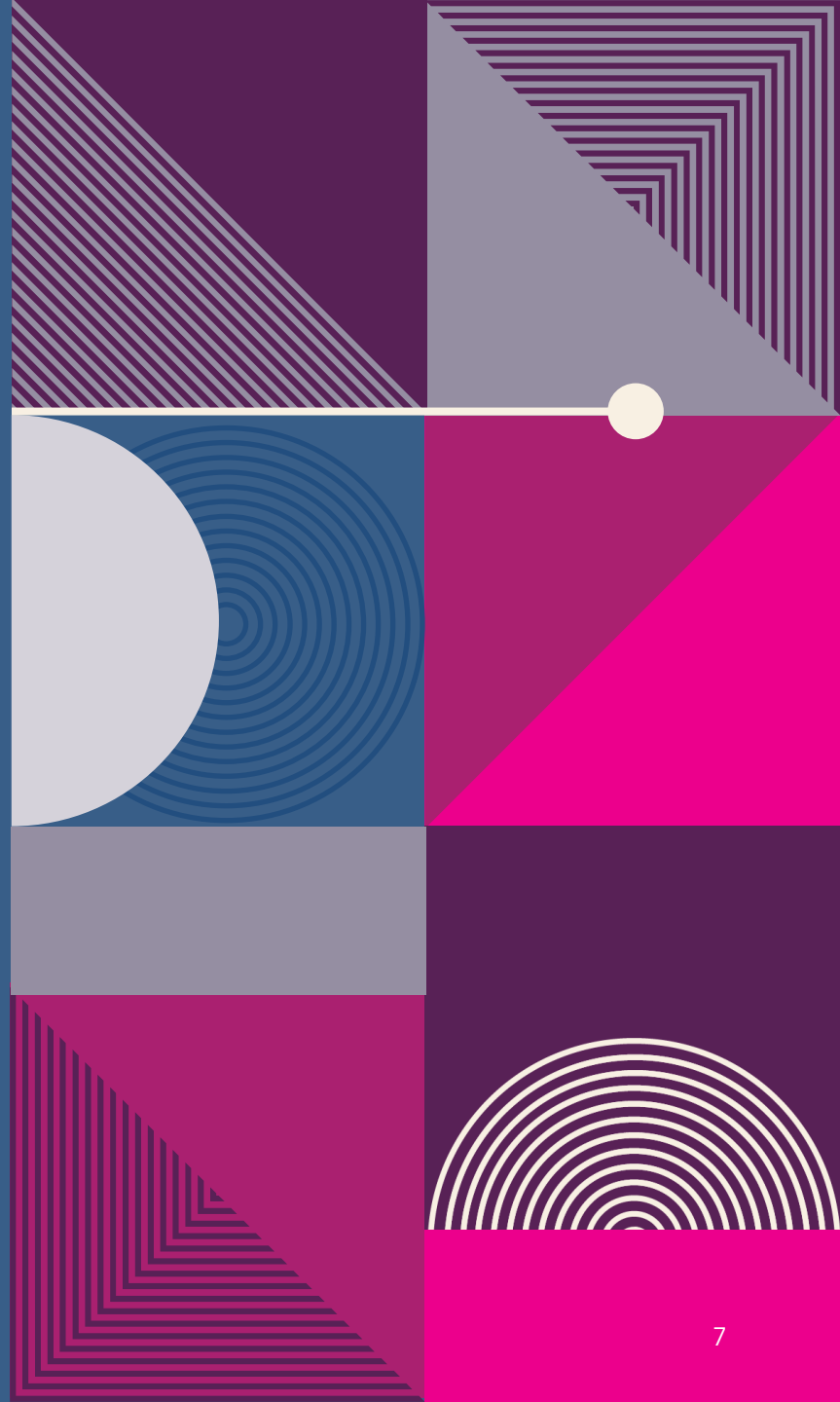
Easily loaded with AutoTokenizer and AutoModelForCausalLM ,
.gninut-enfi dna noitatnemelpmi sefiilpmis hcihw

☑ **Open-source & Actively Maintained**

Supported by the community, ensuring improvements and updates.

TrainingArguments Explanation

- **output_dir="./results_mnad"**
This is the folder where the trained model and checkpoints will be saved.
- **overwrite_output_dir=True**
Allows overwriting the output folder if it already contains files.
- **num_train_epochs=1**
The model will go through the entire training dataset once (1 epoch).



CONTINUE

1.per_device_train_batch_size=4
Number of training samples per batch, per device (GPU or CPU).

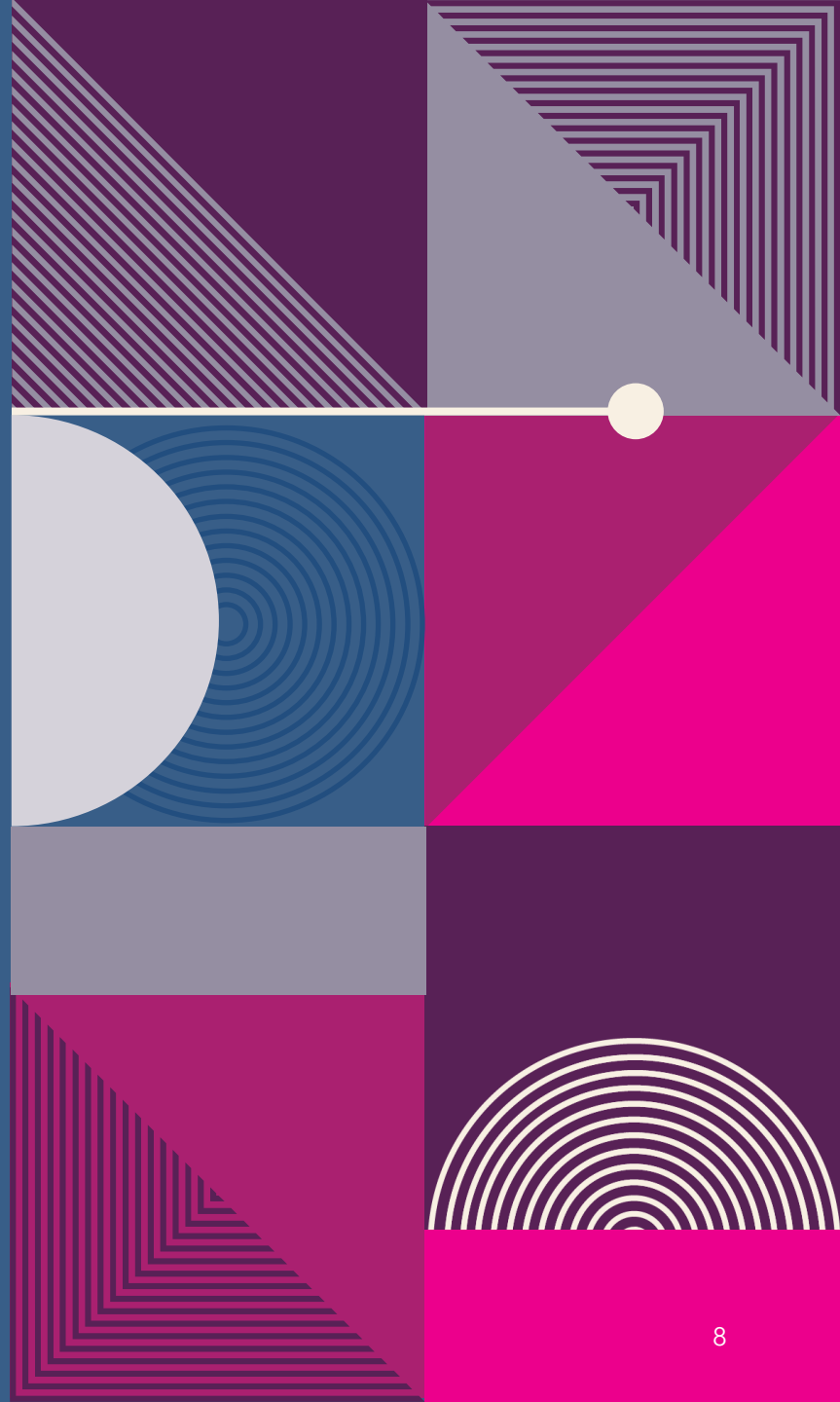
2.save_strategy="epoch"
Saves the model automatically at the end of every training epoch.

3.logging_dir="./logs_mnad"
Directory to store logs (such as loss and accuracy) for visualization (e.g., in TensorBoard).

4.logging_steps=10
Logs training information every 10 steps, such as loss values.

5.fpeurT=16
Enables training with dna ecnamrofrep retsaf rof noisicerp tib-16 .(UPG seriuqer) egasu yromem rewol

6.report_to="none"
Disables reporting to external monitoring tools like WandB or TensorBoard
7.save_total_limit=1
Keeps only the most recent checkpoint to save disk space.



generate_completions() Function

This function generates text completions based on user input using the AraGPT2 mode

- **Input Check**

If the input is empty, it returns a default message repeated num_suggestions times.

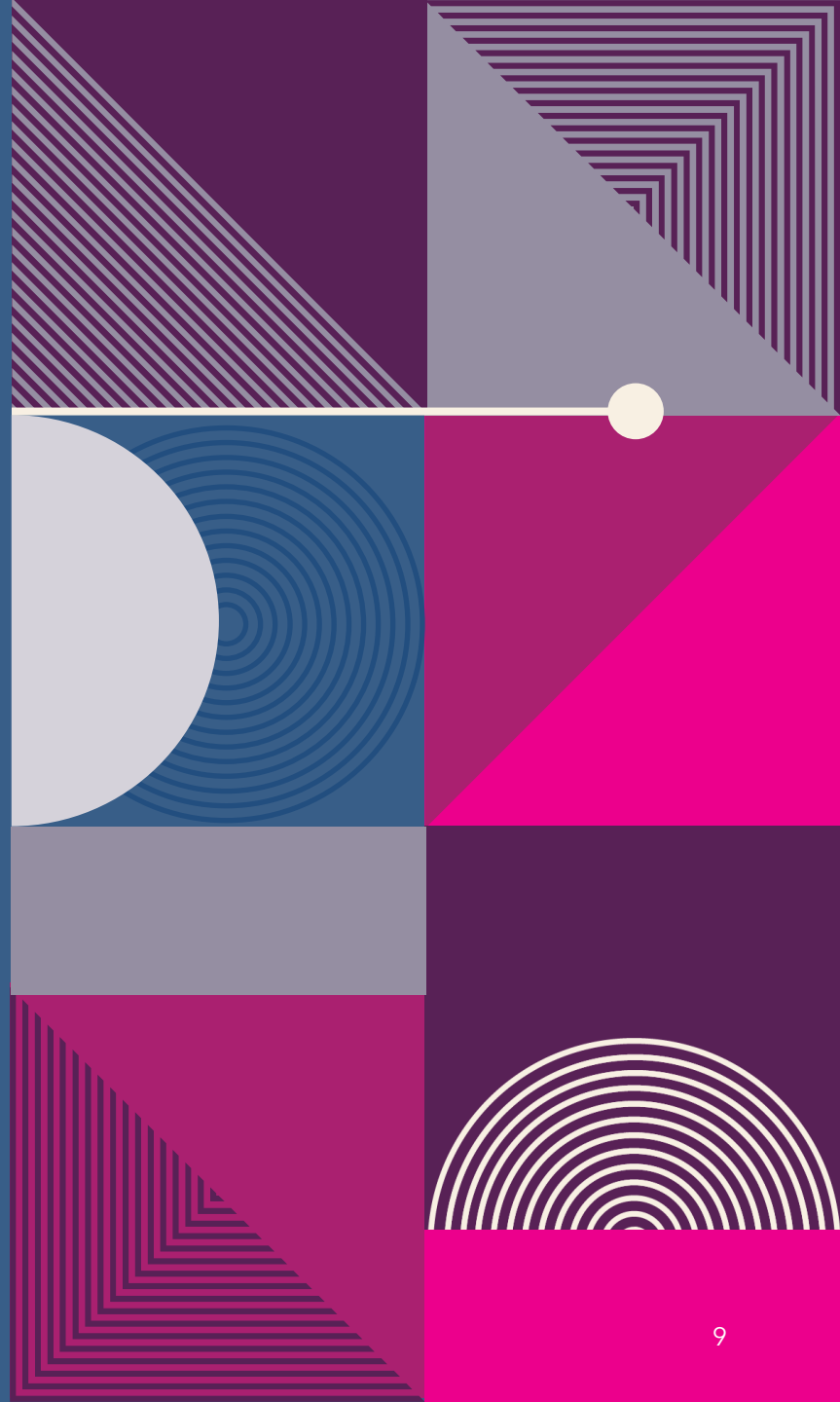
- **Tokenization**

The input text is tokenized using the model's tokenizer and moved to the correct device (CPU/GPU).

- **Text Generation**

The model uses generate() to create multiple text sequences with:

- max_length.htgnel tuptuo latot stimiL :
- num_return_sequences.nruter ot snoitseggus fo rebmuN :
- do_sample=True.gnilpmas modnar selbanE :
- top_p=.ytisrevid rof gnilpmas suelcuN :0.9
- temperature.(evitaerc erom = rehgih) ssenmodnar slortnoC :



EVALUATION PIPELINE OVERVIEW

•We begin by **loading a cleaned Arabic text file** and converting it into a structured pandas DataFrame for processing.

•Next, we **split the data manually** into training and testing sets using train_test_split10% gnipeek ,for evaluation.

•To speed up evaluation, we **sample 1000 examples** from the test set and convert them into a HuggingFace Dataset format.

•The selected text samples are then **tokenized using the AraGPT2 tokenizer** ot deilppa noitacnurt dna gniddap htiw , .shtgnel tupni tnetsisnoc erusne

CONTINUE

1.We define TrainingArguments specifically for evaluation, setting the batch size, disabling logging to external systems, and specifying the output directory.

2.A DataCollatorForLanguageModeling is used to **prepare the input batches for causal language modeling** hcihw , — txetnoc suoiverp no ylno desab drow txen eht stciderp unlike masked language modeling used in BERT.

3.We initialize a HuggingFace Trainer with the model, tokenizer, data collator, and evaluation settings.

4.The trainer runs the **evaluation process** on the tokenized test dataset, calculating the loss.

5.Finally, we compute the **perplexity** using the evaluation loss —a key metric that measures how well the model can predict the next word (lower values indicate better performance).

نظام الإكمال التلقائي للنصوص العربية

اكتب بداية الجملة بالعربية واحصل على اقتراحات ذكية.

النص المدخل

الشباب المغربي

أقصى طول للإكمال

50

10 150

عدد الاقتراحات

3

1 5

درجة الإبداع

0.4

0.1 1

إنشاء الاقتراحات

مسح

اقتراح 1

الشباب المغربي يتعاقد مع لاعب بارز تعاقد فريق شباب الريف الحسيمي لكرة القدم، مع اللاعب عبد الرزاق هيفتي؛ خلال مرحلة الانتقالات الصفية الحالية صلاح مغاني على الساعة h تعاقد نادي شباب الحسيمة لكرة اليد؟ مع المهاجم عبد عبدالرزاق

اقتراح 2

الشباب المغربي يحقق فوزا مهما على مولودية الجزائر حقق فريق شباب الريف الحسيمي فوزا كبيرا على مضيفه مولودية الجزائري، بهدفين دون رد؛ في المباراة التي جمعتهما مساء اليوم الأحد؟ لحساب الجولة الرابعة من دور المجموعات لمسابقة كأس الاتحاد الإفريقي لكرة القدم صلاح

اقتراح 3

الشباب المغربي يواجه نهضة بركان في نصف نهائي كأس العرش يواجه فريق شباب الريف الحسيمي لكرة القدم، اليوم السبت نادي نهضة الزمامرة؛ برسم نصف النهائي من كأس الملك محمد السادس للأندية الأبطال إلياس البطاحي على الساعة h يواجه نادي شباب الحسيمة

THANK YOU

Brita Tamm

502-555-0152

brita@firstupconsultants.com

www.firstupconsultants.com