ARABIC AUTOCOMPLETE SYSTEM

Smart Arabic Word Suggestions in Real Time

NLP Group Project
Team #12

OUR TEAM

1. Mohamed Nasser
2. Mohamed Abdelrazak
3. Mohamed Awad
4. Mohamed Ayman
5. Youssif Nady
6. Osama Samir

INTRODUCTION

This project is a real-time Arabic autocomplete assistant designed to enhance typing experience in Arabic using AI.

WORK FLOW

1. User types Arabic word.
2. Text is normalized and tokenized.
3. Model predicts top 3 next tokens.
4. Suggestions displayed as clickable buttons.
5. Clicking a button appends the word.

PREPROCESSING

- Arabic Normalization :
  - Remove diacritics
  - Remove Tatweel: (ـ)
  - Normalize Alif variants: ا → آ ,ا → أ ,ا → إ
- This ensures more consistent and accurate predictions.

THE MODEL

1. **Model Name:** aubmindlab/aragpt2-base
2. **Architecture:** GPT-2 (Decoder-only transformer)
3. **Parameters:** ~124M
4. **Tokenizer:** Byte-Pair Encoding (BPE) adapted for Arabic
5. **Training Objective:** Causal Language Modeling (predict next token)

DATASET

- Dataset Used by AraGPT2
- Source: Arabic portion of OSCAR and other crawled web text
- Domain: Diverse (news, blogs, forums)
- Size: ~100GB of cleaned Arabic text

MODEL LIMITATIONS

1. May generate grammatically incorrect or nonsensical completions.
2. Struggles with dialectal Arabic or rare words.
3. Only works one word at a time (not full sentence predictions).
4. Not fine-tuned on chat data specifically.
5. Doesn’t handle context across multiple lines.

CONCLUSION

This project highlights the potential of modern NLP models to improve Arabic user experiences across digital platforms.

THANK YOU