SRTTU

Journal of Electrical and Computer Engineering Innovations (JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Using GPT-2 Model and Hazm Library for Persian Text Generation

M. Solouki, Z. Askarinejadamiri [*]*, N. Zanjani*

Computer science of Refah University college, Tehran, Iran.

Article Info

Article History:
Received 10 September 2023
Reviewed 20 October 2023
Revised 18 December 2023
Accepted 25 December 2023

Keywords:
Persian text generation
Hazm library
GPT-2 model

[*]Corresponding Author's Email Address:
Askarinejad@refah.ac.ir

Abstract

Background and Objectives: This article explores a method for generating Persian texts using the GPT-2 language model and the Hazm library. Researchers and writers often require tools that can assist them in the writing process and even think on their behalf in various domains. By leveraging the GPT-2 model, it becomes possible to generate acceptable and creative texts, which increases writing speed and efficiency, thus mitigating the high costs associated with article writing.

Methods: In this research, the GPT-2 model is employed to generate and predict Persian texts. The Hazm library is utilized for natural language processing and automated text generation. The results of this study are evaluated using different datasets and output representations, demonstrating that employing the Hazm library with input data exceeding 1000 yields superior outcomes compared to other text generation methods.

Results: Through extensive experimentation and analysis, the study demonstrates the effectiveness of this combination in generating coherent and contextually appropriate text in the Persian language. The results highlight the potential of leveraging advanced language models and linguistic processing tools for enhancing natural language generation tasks in Persian. The findings of this research contribute to the growing field of Persian language processing and provide valuable insights for researchers and practitioners working on text generation applications in similar languages.

Conclusion: Overall, this study showcases the promising capabilities of the GPT-2 model and Hazm library in Persian text generation, underscoring their potential for future advancements in the field This research serves as a valuable guide and tool for generating Persian texts in the field of research and scientific writing, contributing to cost and time reduction in article writing.

Introduction

Text generation is a field within natural language processing (NLP) that combines computational linguistics and artificial intelligence to automatically create written text to meet specific communication needs. Advances in AI have greatly improved our ability to generate targeted texts for various purposes. The study of text generation dates back to the 1970s, with early work by Golding focusing on natural language generation (NLG) based on deep conceptual understanding. Substantial progress was made in this field during the 1980s.

McDonald viewed text generation as a decision-making problem [1]. Later, the concept of Natural Language Generation (NLG) was introduced, referring to the software process that converts data into human-like language. In 2019, OpenAI released an AI engine called GPT-2, capable of generating text. GPT-2 could process text based on specific guidelines and even had an experimental online tool called Talk To Transformer. This

Doi: 10.22061/jecei.2023.10101.690

AI engine possessed powerful text generation and prediction capabilities, particularly suited for smartphones. To improve GPT-2, a machine learning system was trained using around 8 million web pages [2]. This system can identify missing words in text and replace them appropriately.

Text production is essential in any language, and text generation serves various applications. GPT-2 allows or assists writers and typists by completing sentences and generating ideas across different domains. GPT-2 has been used for text generation in multiple languages, including English, German, and Bengali.

GPT-2's language modeling algorithm demonstrates coherent semantic capabilities and possesses a unique generative feature found in human language. Language generation involves producing new linguistic expressions based on existing language rules, and GPT-2 excels in this aspect [3]. The training of this language model involved utilizing over 40 gigabytes of web data and nearly 1.5 billion parameters of text structures.

GPT-2 exhibits a wide range of capabilities, such as generating high-quality conditional text samples when given input. It outperforms other language models in specific domains like Wikipedia, news, or books without the need for extensive training datasets [4], [5]. GPT-2 can learn. In tasks such as comprehension, summarization, and translation, GPT-2 learns these tasks from raw text without relying on specific training data. By using GPT-2, we can facilitate the writing process, reduce energy consumption, increase speed, and improve efficiency. For instance, when writing abstracts or articles that would otherwise take hours, GPT-2 helps reduce time, cost, and energy consumption, highlighting the necessity and importance of this field. GPT-2 also contributes to cost reduction in scientific production, which can be a costly endeavor [6], [7]. A similar example is the auto-complete feature when individuals search on Google. It helps users reach their intended search queries faster by suggesting word completions, reducing costs and time spent on multiple searches.

Guessing the next word, which may occur billions of times, has enabled this AI-based system to produce text that appears to be written by a human. The generated texts from this text processing engine are grammatically correct, have a consistent and coherent theme, and possess an engaging style of expression. Therefore, intelligent text generation has been performed in different languages using specific algorithms and artificial intelligence algorithms. In this research, we aim to enable text generation in the Persian language.

Related Work

A. Language Modelling

Using a starting word to create meaningful sentences is crucial in natural language processing. The question of whether machines can think and be creative like humans is what we aim to answer in sentence construction. We train systems for specific tasks and apply them in natural language processing to tackle challenges in sentence generation, such as text summarization, machine translation, and automated question answering.

Language modeling (LM) involves predicting the next word in a given sequence of data. Researchers have recently taken a keen interest in LM in natural language processing. Language models can be divided into count-based models and continuous space models [8].

Count-based models use statistical formulas to describe the language model and construct the joint probability distribution of word sequences. An example is the n-gram model, which predicts one word at a time based on the Markov assumption. The probability of a word sequence is calculated as the product of word probabilities based on previous words, with a history limited to a certain number of words.

Continuous space models include Neural Language Models (NLMs). There are two main types of NLMs: feed-forward neural networks and recurrent neural networks. Feed-forward neural networks address data sparsity, while recurrent neural networks overcome the limitation of context. Recurrent neural networks have shown advanced performance. Feed-forward neural networks have hierarchical probabilistic models and HLBL models that improve training speed.

Various methods exist for text generation, including Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) models, encoder-decoder frameworks, and sequence-to-sequence models. RNNs and LSTMs generate sentences starting from a given word, while encoder-decoder models encode sentences into fixed-length vectors and then use the vectors to generate sentences. However, Transformer models like OpenAI GPT-2 and BERT have also proven effective in text generation.

GPT-2 language models consist of three components. "Generative" means the model is trained to predict or generate the next element in a sequence, even non-sequentially [9]. Prediction is a crucial task in scientific and industrial applications. GPT-2 can generate diverse and meaningful text when provided with raw textual data, without human intervention. It excels in language tasks like machine translation. GPT-2 is based on the Transformer architecture, which is known for language modeling, translation, and classification tasks, as well as computational efficiency [10]. Transformers differ from previous technologies like RNNs and LSTMs by not relying on previous states, resulting in faster computation. Machine translation involves converting a sentence from one language to another, and Transformers, with their encoder and decoder sections, facilitate this process. GPT-2 has been widely used for text generation in various languages, including Chinese, Japanese, German, and Persian (referred to as "Bolbol-e-Zaban" in the presented work for generating poems and verses using AI).

B. Reviewing Text Generation in Different Language

In Chinese text generation, words are masked in sentences, and the BERT model predicts the masked labels [11]. GPT-2 is built with a Transformer decoder module that takes a starting token as input and generates tokens one at a time, creating the sequence. BERT, on the other hand, uses a bidirectional encoder based on the Transformer and predicts masked labels using its Masked Language Model (MLM) task. Generating sentences based on starting words is the main task when using BERT.GPT-2, an advanced transformer-based model, was used to generate artificial text samples by training on various user prompts as input. The original version had 1.5 gigabytes of parameters, and a smaller version with 117 megabytes of parameters was later released for fine-tuning on custom text datasets. In this particular language generation task, Jpop text was used as the training data, with each text delimited by <I endoftext I>.

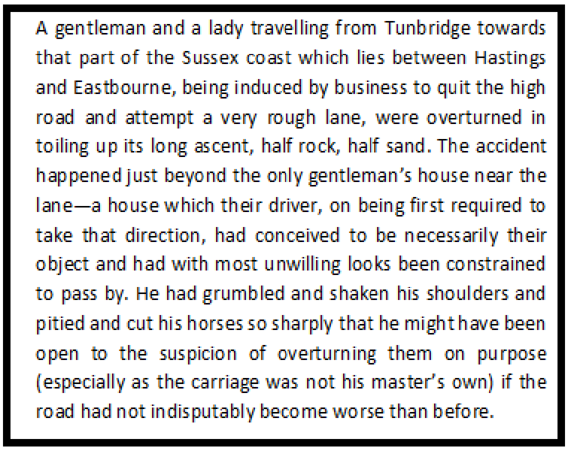
In paper under title of “Toward Russian Text Generation Problem Using OpenAI’s GPT-2” discuss around text generation problem in Russian language. This work focuses on Natural Language Generation (NLG) and explores modern approaches using deep neural networks. Specifically, it examines popular NLG solutions based on the Transformers architecture with pre-trained models like GPT-2 and BERT. The challenge lies in the limited availability of Russian language models that can generate text within specific subject areas. The objective of this study is to develop a model capable of generating contextually coherent narrow-profile text in Russian. As part of the study, a model was trained to generate coherent articles in Russian within a specific subject area, along with a software application for interacting with it [12]-[14]. In this study, we try do same process to generate Persian text [12].

The field of question answering (QA) and text generation using BERT and GPT-2 transformers is a specialized area within information retrieval systems. It involves formulating queries in natural language and retrieving the most accurate or closest answer. The primary goal of QA systems is to provide concise answers to questions, rather than presenting a list of relevant documents. On the other hand, text generation focuses on producing coherent English text by predicting the next sentence or generating content based on previous words [15], [16].

The latest version (3.1.0) of Huggingface transformers was used to generate German text. The new Trainer class was utilized, and the GPT-2 model was fine-tuned using German recipes from chefkoch.de. German text data for GPT-2 training was obtained from the Huggingface model hub, specifically from German recipe datasets with metadata from chefkoch.de [11]. The aim was to fine-tune GPT-2 using recipe descriptions and generate new recipes [14], [17], [18].

In another study, researcher focus t generate Sepedi text. This study focuses on developing and analyzing a language model called Generative Pre-Trained Transformer 2 (GPT-2) to generate phrases in the Sepedi language. Sepedi is a language that lacks resources and poses challenges due to its orthographic representation. The GPT-2 transformer requires large datasets and advanced computational resources. The researchers used the National Centre for Human Language Technology (NCHLT) Sepedi text dataset, which is unstructured. Despite working with a small dataset, the developed text generation model achieved a low loss value of 2.36. The generated text produced by this model is syntactically correct but contains some grammatical errors. Overall, the model outperformed previous Sepedi text generation models that used transformer-based techniques [19]-[21].

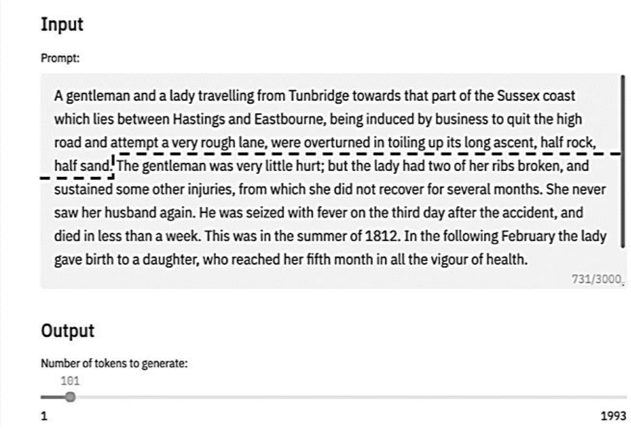
Recently, The Guardian published an article about GPT-3, stating that it can write a 500-word article using simple and clear language especially in English. The article emphasized that there is no need to fear artificial intelligence. The potential impact of machine learning across various fields, such as medicine, law, sociology, and communications, in the coming decade is expected to be unprecedented [22], [23]. GPT-3 generates text that is statistically appropriate. For example, the initial description of a car accident resembles the opening sentence of Jane Austen's Sanditon. This is the latest achievement of GPT-3, as described in the original text (Fig. 1 & Fig. 2).



A gentleman and a lady travelling from Tunbridge towards that part of the Sussex coast which lies between Hastings and Eastbourne, being induced by business to quit the high road and attempt a very rough lane, were overturned in toiling up its long ascent, half rock, half sand. The accident happened just beyond the only gentleman’s house near the lane—a house which their driver, on being first required to take that direction, had conceived to be necessarily their object and had with most unwilling looks been constrained to pass by. He had grumbled and shaken his shoulders and pitied and cut his horses so sharply that he might have been open to the suspicion of overturning them on purpose (especially as the carriage was not his master’s own) if the road had not indisputably become worse than before.

Fig. 1: text generation of description of a car accident.

We provided GPT-3 with the first sentence and compared the previous text draft and text generation using GPT-2 by referring to Fig. 1 and Fig. 2. Both GPT-3 and GPT-2 function like search engines, where the generated text by GPT-3 is an additional line. Similar to how Google provides relevant answers to our queries without understanding them, GPT-2 follows the given text and seamlessly continues the sequence of our words, extending it to a specified length [24]-[26].



Input

Prompt:

A gentleman and a lady travelling from Tunbridge towards that part of the Sussex coast which lies between Hastings and Eastbourne, being induced by business to quit the high road and attempt a very rough lane, were overturned in toiling up its long ascent, half rock, half sand. The gentleman was very little hurt; but the lady had two of her ribs broken, and sustained some other injuries, from which she did not recover for several months. She never saw her husband again. He was seized with fever on the third day after the accident, and died in less than a week. This was in the summer of 1812. In the following February the lady gave birth to a daughter, who reached her fifth month in all the vigour of health.

731/3000

Output

Number of tokens to generate: 101

1

1993

Fig. 2: Text generated by gpt-2.

This study involved focus of GPT-2 in generating text across various languages. Through extensive experimentation and evaluation, it was observed that larger datasets produced more accurate and meaningful samples compared to smaller ones. Moreover, the application of specific language-specific libraries, such as the Hazm library for Persian, significantly improved the quality of generated text, ensuring correct spelling and adherence to linguistic rules. The study successfully demonstrated the capability of GPT-2 in generating content-rich text in Persian, highlighting its potential as a valuable tool for text generation in the Persian language. With the advancements in language models like GPT-2, the field of text generation is continuously evolving and promising, providing new possibilities for researchers and language enthusiasts alike.

C. Language Model in Different Languages

Having explored text generation in different languages and discussed various language models, let's now delve into the discussion of text generation using different decoding methods.

Several algorithms are used for text generation, including [27]-[30]:

1. Greedy search: Choosing the next word based on the highest probability, which is the most probable word.

2. Beam search: Maintaining a set number of hypotheses at each step and selecting the hypothesis with the highest probability. This method reduces the risk of losing highly probable word sequences. Transformers can utilize beam search by setting parameters like num_beams > 1 and early_stopping = True to stop generation when all beam hypotheses reach the end of sentence (EOS) token.

Recently, beam search has been criticized for not being the optimal choice due to the following reasons:

- Beam search works well when the desired generation length is predictable, such as in machine translation or text summarization tasks.
- Beam search struggles with generating repetitive text, and controlling it through measures like n-grams or penalties is challenging.

To address these limitations, Holtzman et al. proposed an alternative method known as sampling. In its simplest form, sampling involves randomly selecting the next word based on its conditional probability distribution. One sampling scheme called Top-K sampling was introduced, where the next words are filtered based on the top-K highest probabilities, and the probability mass is redistributed only among those K words. GPT-2 adopted this sampling scheme, contributing to its success in generating stories.

The word domain used in the example above was expanded from 3 words to 10 words to better illustrate Top-K sampling. One issue with traditional sampling is that it lacks dynamic adaptability to the number of words filtered from the probability distribution of the next word (P(w | w1:t-1)). To address this, Top-p sampling is employed. Instead of sampling solely from the top-K probable words, Top-p sampling selects from a smaller set of words whose cumulative probability exceeds a certain threshold, denoted as p. The probability mass is then distributed among this set of words. This approach allows the size of the word set to dynamically change based on the probability distribution of the next word [9].

In general, when using a text generation algorithm, it is important to first determine the objectives of the text generation task and then experiment with different algorithms to select the most suitable one. In this study, various language models were explored in different languages, including English, German, Chinese, Russian and Sepedi. Through extensive analysis, it was evident that GPT-2 stands out as a highly effective language model for text generation in Persian as well. By fine-tuning GPT-2 with the Hazm library, the generated Persian text exhibited coherence, accuracy, and adherence to linguistic rules. The successful application of GPT-2 in Persian text generation opens new horizons for researchers, writers, and content creators in the Persian language domain. The integration of language-specific libraries, like Hazm, proved crucial in achieving superior results, ensuring that the generated text reflects the linguistic nuances and intricacies unique to Persian. Overall, the study's findings demonstrate the immense potential of GPT-2 in generating high-quality and contextually relevant text in Persian, showcasing its versatility as a language model for diverse linguistic applications.

Proposed Method

For this study, our goal was to explore relevant articles on text generation from resources like Google Scholar, ScienceDirect, and websites such as GitHub, Hugging Face, and Medium. The aim was to find existing libraries for text generation in the Persian language. To achieve this, we utilized the GPT-2 tool and online resources like Google Colab. We also made use of libraries such as regex, Transformer, and Hazm. Text generation was performed using the GPT-2 tool, where sentences were initiated with a starting word and tokenization was applied using GPT-2 and various libraries and tools.

As machine learning algorithms require input data, it was necessary to prepare a dataset. Google Colab, a free cloud service provided by Google, was used for implementation and text generation. It allows programming in Python and offers the capability to install and work with various Python packages. The proposed method of this paper combines various deep learning libraries and text processing tools to build and evaluate language models for text generation in the Persian language.

1. Pipeline: The method utilizes the Pipeline feature available in libraries like Hugging Face's Transformers. The Pipeline offers a convenient API for various natural language processing tasks, including text generation. It allows for easy integration of pre-trained language models and tokenizers, simplifying the process of building language models for specific tasks.

2. PyTorch: To build one of the language models, the proposed method employs PyTorch, a flexible and powerful deep learning library. PyTorch enables the creation and training of custom neural network architectures tailored for Persian text generation. The model is trained on a sizable dataset of Persian text using PyTorch's efficient optimization and GPU capabilities.

3. Tokenizer: The proposed method incorporates tokenization libraries compatible with PyTorch and other deep learning frameworks. Tokenizers are essential for breaking down text into smaller units and encoding them as input for language models. These libraries facilitate efficient tokenization and data preparation for training the models.

4. Regex: The method also utilizes regular expressions (regex) for text preprocessing tasks. Regex is valuable for pattern matching and replacing specific strings, allowing for data cleaning and normalization before feeding the text into the language models. This step ensures that the input data is in the appropriate format for training.

5. Transformer: The proposed method focuses on using the Transformer architecture for language modeling. Transformers have proven to be highly effective in natural language processing tasks, including text generation. The Transformer's self-attention mechanism enables the model to capture contextual dependencies effectively, leading to more coherent and contextually relevant text generation.

6. Hazm Library: For language-specific preprocessing and normalization tasks in Persian, the Hazm library is utilized. Hazm offers features such as sentence and word segmentation, stemming, and handling half-space characters, specifically designed for Persian text processing. The integration of the Hazm library ensures that the language model can handle the unique linguistic characteristics of Persian.

By combining the Pipeline, PyTorch, tokenizer libraries, regex, Transformer architecture, and the Hazm library, the proposed method aims to create comprehensive language models for Persian text generation. This multi-tool approach will enable a thorough evaluation of the models' performance and their effectiveness in generating coherent and contextually appropriate text in the Persian language. The paper's findings will provide valuable insights into the strengths and limitations of each tool for text generation tasks in Persian, guiding researchers and developers in building powerful language models for Persian natural language processing.

Average consumed resources were calculated in the middle of Google Colaboratory after many launches of each of the original models. The results can be seen in Table 1.

Table1: consume source

Number of articles	Occupied disk space, GB	RAM, GB	GPU memory, GB
Gpt-2 100	0.7	2.63	2.52
Gpt-2 500	1.32	3.56	4.63
Gpt-2 1000	4.5	6.53	8.3
Gpt-2 5000	5.3	8.34	9.33
Gpt-2 20000	8.26	11.3	12.5

Implementation Results and Discussion

Generating Persian text with artificial intelligence tools is feasible, and this section examines models with different datasets and sizes. In this study, the execution method involves removing HTML tags using the Beautiful Soap library from a dataset of 100 articles, leaving only the Persian text. Next, the data is normalized using the Hazm library. Certain characters in the dataset are replaced using the regex library, such as "ی" and "ک". Tags like <Num> and <LAT> are used to represent English letters and numbers. The <eos> tag indicates the end of sentences, while the [STARTABS] and [ENDABS] tags mark the beginning and end of phrases to improve data structure.

Model Training

After installing the required libraries (transformer, tokenizer, pipeline, etc.), the model is retrained. It is then trained with different initial inputs, such as "پژوهشگران", "در این کار تجربی ابتدا" and "مطالعه برخی خواص to generate samples and perform text generation (shown in Table 2 and Table 3). There are also several launches of all models with different input phrases such as "پژوهشگران "،"دراین "مطالعه برخی خواص" and" کارتجربی ابتدا".

Table 1: comparing beam search and to-k search result

Beam search	Top-k search
آب مروارید کودکان شایع می باشد. گروه آزمایشی در دو گروه آزمایشی و کنترل، مکمل ویتامین دریافت کردند و گروه کنترل نیز مکمل ویتامین دریافت کردند. گروه آزمایشی در دو گروه آزمایشی و کنترل، مکمل ویتامین دریافت کردند و گروه کنترل نیز مکمل ویتامین دریافت کردند. در این مطالعه مقطعی، از کودکان مبتلا به در دو گروه آزمایشی و کنترل، خون گیری به عمل آمد. کودکان گروه آزمایشی در دو گروه آزمایش و کنترل، به طور تصادفی به دو گروه آزمایش و کنترل تقسیم شدند. کودکان گروه آزمایشی در دو گروه آزمایش و کنترل، به طور تصادفی به دو گروه آزمایش و کنترل تقسیم شدند	آب مروارید کودکان با روش توصیفی تحلیل کوواریانس از دید ساله شهر Nکودکان زیر تبریز را مورد بررسی قرار داده اند. جمع آوری داده ها بر اساس روش توصیف کوواریانس در دو مرحله تجزیه شده و با استفاده از نرم افزار و آزمون های تی تست انجام گردید. یافته ها: نتایج نشان داد که بین سبک های دلبستگی والدین و هوش هیجانی کودکان رابطه مثبت و معنادار وجود دارد؛ اما در سطح معناداری هوش هیجانی، هوش عاطفی رابطه معنی دار را با هوش هیجانی کودکان برقرار نمی کند. نتیجه گیری: ها بیانگر نقش یافته میانجی گری والدین و هوش هیجانی است و می تواند حاکی از نقش والدین در رابطه عاطفی بهتر کودکان باشد.به منظور تعیین ارزش ویژه هوش هیجانی در کودکان تیزهوش، نتایج یک پژوهش توصیفی تحلیلی شهر تبریز انجام بر روی شد. جامعه آماری پژوهش، کودکان تیزهوش و کم شنوای منطقه۳ ساله شهر تبریز بودند.

In order to compare suitable algorithm, in one example, the beam search algorithm with the num_beams parameter did not produce satisfactory results, as it led to word repetitions in the generated samples (Table 1). Subsequently, the top-k and top-p algorithms were used, resulting in samples with lower error rates. Table 1 demonstrates the output obtained using the top-k algorithm.

Initially, a 500 parameter-model was evaluated with different inputs, and the results were analyzed. The 500-model, when given that input produced the samples shown in Table 3. The section compares the 500-model with and without the Hazm library applied. The samples indicate that the 500-model with the Hazm library produces more accurate sentences in terms of grammar and appearance (Table3).

Table3: Comparing Model with and without Hazm

Start word	Model without Hazm	Model with Hazm
مطالعه برخی خواص500	مطالعه برخی خواص شربت سنجد و سنجد سنجد سنجد و سنجد سنجد	مطالعه برخی خواص منحصر بفرد گل میخک می توانید از آن به عنوان یک ماده ضد التهاب و ضد سرطان نام ببرید . بررسی ها مشخص کردند میخک منبع خوبی از آنتی اکسیدان به عنوان یک ضد التهاب قوی است که التهاب را کاهش می دهد و از ابتلا به پلاک های خونی جلوگیری می کند. همچنین از آن به عنوان یک گیاه ضد انعقاد خون ، نوعی آنتی بیوتیک طبیعی نام می برند. میخک یکی از موثر ترین و مفید ترین داروهای طبیعی شناخته شده است و همچنین از آن به عنوان یک ماده ضد حساسیت و کاهش دهنده التهابات بدن نام برده میشود. این گیاه یکی از بهترین عصاره های گیاهی شناخته شده در درمان آلرژی های غذایی و اگزما میباشد . ماده ای که در این گیاه می توان یافت .
پژوهشگران 500	پژوهشگران با بررسی و مطالعه بر روی تعداد مشخصی از میکروب ها موفق شدند به نتایج جالب توجهی برسند . در حالی ه با بررسی میکروبی به نام شoOه و مقایسه باکتری هایG ه و میکروسینیتی دریافتندe ه به طور مرتب آب و مایعات بیشتری در دهان خود میبلعد. به گفته کارشناسان میکروب شbه در برابر B ه شXه با قرار گرفتن در معرض ه آب کمتری مینوشد. در حالیm ه آب بیشتری در دهان خود میسوزاند ه ش٪ه آب بیشتری در دهان میبلعد.	پژوهشگران با استفاده از تکنیک هایی خاص به بررسی رفتار ژنتیکی باکتری های روده پرداختند. این باکتری ها به شکل خاصی نسبت به مواد آنتی باکتریایی واکنش نشان می دهند و به طور معمول باعث بروز علائم مسمومیت می شوند. آزمایش آنها توانست با انجام مطالعاتی بر روی باکتری های روده باعث بررسی دقیق آن بیماری می شود. در طی این تحقیق باکتری های روده ای مورد بررسی قرار گرفتند. نتایج این بررسی در نشریه های معتبر علوم مولکولی دنیا منتشر شد و به عنوان یکی از مقالات مهمی که با موفقیت چاپ گردید، در نظر گرفته شد. نتایج این تحقیق نشان داد باکتری های روده ای نسبت به مواد آنتی باکتریالی مقاومند. هدف از این مطالعه اندازه گیری آنتی باکتریال و آنزیمی بود که در باکتری های روده وجود دارد. میکروب های روده در طی سالهای متوالی نسبت به این مواد حساسیت نشان داده و به باکتری های گرم منفی واکنش نشان داده و موجب بروز علایم مسمومیت میشوند. اکثر باکتری های روده مقاوم هستند. لذا بر کاهش ویا دفع باکتری های مفید روده اثر میگذارند. هدف از این تحقیق مطالعه تأثیر میکروبیتاسیون بر باکتری های روده بود.

Next, the 1000 parameter -model was considered in two scenarios: with and without the Hazm library. The model was tested with different inputs. Ordered and disordered English and Persian letters in the generated sample, leading to some words and sentences being nonsensical. However, this issue is resolved in the subsequent example by applying the Hazm library. The 1000-model, when not normalized and without the Hazm library, produced nonsensical words and fragmented English words in the samples. In contrast, the 1000-model with the Hazm library applied generated samples with correct spelling and linguistic aspects. It's important to note that the generated samples effectively convey their meanings. They consist of coherent and relevant words and sentences related to the discussed topics, indicating accurate content. For example, we present a sample from the10000 and 50,000 parameter-model. We then compare the 10,000_Hazm model with the Bolbolzaban model as a persian text generator model. Table 4 illustrate the output of the Bolbolzaban model for generating text with the different inputs (Table 4). Initially, the "Bolbolzaban" model was used for training the data, and the generated samples from the Bolbolzaban model were compared to the 10,000-case data model as described earlier.

Table 4: comparing Hazm Model with Bolbolzaban

	Hazm	Bolbolzaban
دراین کارتجربی 10000	دراین کارتجربی ، لایه های نازک اکسید مس- و روی در آب تهیه شده است. در این مقاله ، بررسی مراحل انجام فرایند جوشکاری سیمی، در قالب طرح توجیهی و عملی ارایه شده است. این مطالعه به منظور بررسی ویژگیها و مراحل مختلف مراحل جوشکاری سیمی در قالب طرح توجیهی و عملی ارائه شده است تا میزان استفاده از فلزات رایج و متداول در فرایند جوشکاری را مورد اندازه گیری قرار داده و مزایای استفاده از این روش در صنایع مختلف مورد بررسی قرار گیرد. به طور کلی، طی این مطالعه، عوامل مختلفی که باعث تسریع و توسعه فرایند جوزایی در صنایع مختلف می گردد، مورد بررسی قرار گرفت و در نهایت در صورت نیاز، روش جوشکاری اکسید مس با روش جوشکاری القایی همراه شده است. با توجه به اینکه اجرای این روش به روش الکترودینامیک محدود بوده و هزینه بسیار بالایی را نیز در بر دارد، تحقیقات لازم در رابطه با پارامترهایی نظیر طول و عمق، مقاومت الکتریکی و طول در جوشکاری انجام شد. روش های مختلف مختلف جوشکاری از قبیل پیچها، جوشکاری لایه ای و برشی در این مطالعه مورد بررسی قرار گرفت که نتایج آن در این مقاله ارایه گردیده است.	دراین کارتجربی، لایه های نازک است تا نگردد پاره ، این جان پرست

The 10,000_Hazm model shows better results compared to the Bolbolzaban model, as depicted in Table 4. The 10,000_Hazm model generates more meaningful samples and longer sentences.

Table 5 the output of the 50,000_Hazm model with the initial inputs "مطالعه برخی خواص", indicating the generation of suitable text.

Table 5: 50000 parameters with Hazm Model

Start Word	Output
مطالعه برخی خواص	مطالعه برخی خواص روغن خردل بر روی پوست صورت، مطالعات زیادی در حوزه سلامت پوستی انجام شده است. هدف از این مطالعه بررسی تاثیر ترکیبات موجود در ترکیب روغن خردل بر روی پوست صورت با استفاده از سه مدل تجربی و انجام چند آزمایش به روش کارآزمایی بالینی با ترکیبی استریل ترکیبی و بدون روغن خردل می باشد. پژوهش حاضر به منظور بررسی ارتباط اسیدهای چرب و عصاره گیاه خردل و میزان بروز آکنه در دو گروه آزمایشی و کنترل دانشجو دانشگاه آزاد واحد کرمانشاه صورت پذیرفت. مواد و روش ها: در این مطالعه با استفاده از آنالیز واریانس، از گروه تجربی استفاده شد. از روش کارآزمایی بالینی، به منظور تعیین میزان بروز آکنه صورت در دو گروه آزمایشی و کنترل، سه گروه شاهد به طور تصادفی، نمونه های پوست افراد گروه تجربی در سالن ورزشی با اسید nمیلی گرم انجام گرفت. آزمایش ها: نتایج حاصل از آنالیز واریانس به عنوان نمونه ها برای آزمون هادر قالب طرح کاملا تصادفی در دو گروه نمونه و شاهد با استفاده از آزمون های استاندارد انجام گرفته است.
پژوهشگران	پژوهشگران توانستند برای نخستین بار اثر ضدقارچی های گیاهی را بر روی سطح برگ از نظر خواص مکانیکی بررسی کنند. از آن به بعد، در هر آزمایش علاوه بر ویژگی های بیوشیمیایی آن، میزان عصاره به کار رفته برای مبارزه با میکروب های عامل بیماری زا بررسی شد. نتایج این تحقیق نشان داد عصاره های گیاهی تولید شده بر سطح برگ درختان لیموترش در مقایسه با عصاره بر روی برگ گیاهان دیگر با درصد تتراکریت کمتر، قابل استفاده بوده است. در این تحقیق اثر ضدقارچ های گیاهی برای گیاه لیموترش موردبررسی قرار گرفت. میزان عصاره مورد بررسی از مخلوط عصاره برگ و پودر عصاره مرکبات و آب لیموترش به دست آمد. پس از گذشت سه ماه، نتایج این آزمایش نشان داد عصاره گیری با عصاره های گیاهی امکان پذیر نبوده است. هم چنین غلظت عصاره های تولیدی از نظر مقاومت به قارچ زایی، نسبت به عصاره گیری به روش های شیمیایی غیر مجاز بیشتر بوده است. نتایج این پژوهش با مقایسه عصاره گیری شیمیایی و خالص، حاکی از آن است که عصاره گیری شیمیایی با عصاره گیری از اسانس لیموترش می تواند باعث کاهش اندوکرین و در نتیجه کاهش مقاومت به این باکتری شود.

The pre-trained model works reasonably well, generating grammatically correct texts while maintaining context. The 50000 of parameter model is expected to have more coherent text than the 20000 of parameter model, but to run a larger model, more resources are needed, and they work longer. For the model for correct work it’s required to train it on a sufficiently large amount of text.

Thus, the primary task before training the model was the search for Persian-language resources with a large database of articles on different topics. Due to the number of articles and the approximate amount of text in them,

Thus, the samples generated using the libraries introduced in this study demonstrate that GPT-2, in addition to text generation and prediction, can also complete text continuations and generate content-rich text. It can be a valuable tool for researchers in providing article abstracts.

Experiments to Explore Model Training

After a series of experiments, studies were carried out on various devices on the basis of which the training took place such as GPU, CPU, TPU. It was decided to configure the server with a connection to Google Colud GPU. The chosen subject area is assistance to researcher and copywriters.

Train The Model to Generate Whole Texts

When the articles become longer, the coherence of the text within them increases. An example can be seen in the figure below with 50,000 parameters.

Table 6: 50000 parameters with Hazm Model

Start Word	Output
پژوهشگران	پژوهشگران ژاپنی تصمیم گرفتند با استفاده از یک روش جدید ریاضی ، هندسه دانان را در پی پاسخ این سوال برانگیزند که میزان دقت دستگاه اعداد صفر و یک در مقایسه با دقت دستگاه اعداد طبیعی چگونه اندازه گیری می شود ؟ آیا آن ها با اندازه گیری این ارقام می توانند عملکرد ریاضی دستگاه اعداد طبیعی را بهبود بخشند؟ شیوه جدید ریاضی این دو ریاضیدان ژاپنی نتیجه آزمایش های اخیر آن ها درباره اندازه گیری دقت دستگاه اعداد طبیعی به وسیله دستگاه اعداد طبیعی است روشی که مدت هاست پژوهشگران علوم محاسباتی ، ریاضی دان ها و فیلسوفان و فلاسفه سراسر جهان در پی آن ها به جستجو و بررسی و کاوش در تئوری های حل مسایل ریاضی درباره اندازه گیری دقت دستگاه های اعداد طبیعی پرداخته اند و تلاش های آنان تاکنون بی نتیجه مانده است بررسی کارایی دستگاه های اعداد طبیعی در ارزیابی و سنجش دقت دستگاه های اعداد طبیعی نتایجی را در پی داشت که در پی آن نقش روش اعداد طبیعی در حل مسایل ریاضی آشکار شده و در واقع روش یادگیری برای کودکان و نوجوانان امروزی تضمین گردیبی است که به آن ها کمک می کند تا درک خود از مسایل ریاضی را نسبت به مسایل تجربی افزایش دهند روش یادگیری دانش آموزان امروزی به دو عامل اصلی « سرعت» و « دقت» وابسته است.
پژوهشگران	

Conclusion

In conclusion, this study focused on generating Persian text using GPT-2, a language model based on the Transformer architecture. Various research studies on text generation in different languages were explored to assess the success rates of machine learning algorithms. Larger datasets yielded more accurate, meaningful, and grammatically correct samples compared to smaller datasets. The system reads and learns from input sentences and data, generating text that is not merely a copy of the content. Libraries like Hazm and regex were used for tasks such as HTML tag removal, data normalization, and string pattern matching. Other libraries like Transformer and tokenizer were also utilized. Different algorithms and parameters were experimented with to improve results. The created models outperformed the bolbolzaban model in text generation. The 50000 model served as the parent model. The results showed that the generated text had the highest correlation with the given words and phrases, resulting in coherent and content-rich texts with correct spelling and adherence to grammatical rules. GPT-2 and GPT-3 technologies have the potential to significantly impact future jobs, streamlining workflows through artificial intelligence. Although these models have their strengths and weaknesses, such as input size limitations and text generation constraints, advancements in language models are expected to address these challenges.

Author Contributions

Z. Askarinejadamiri: Supervision, Project administration, Conceptualization, Methodology, Visualization, Investigation, Writing Reviewing and Editing, Programmer, Writing - Original draft preparation.

M. Solouki: Programmer, Validation, Conceptualization, validation, Investigation, collected the dataset, Writing-Reviewing and Editing, Writing-Original draft preparation.

N. Zanjani: Writing Reviewing and Editing. All authors discussed the results.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

GPT	Generative pre-trained transformer
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Networks

References

[1] D. McDonald, P. Proctor, W. Gill, S. Heaven, J. Marr, J. Young, "Increasing early childhood educators' use of communication-facilitating and language-modelling strategies: Brief speech and language therapy training," Child Lang. Teach. Ther., 31(3): 305-322, 2015.

[2] E. S. Jo, T. Gebru, "Lessons from archives: Strategies for collecting sociocultural data in machine learning," in Proc. the 2020 Conference on Fairness, Accountability and Transparency: 306-316 2020.

[3] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," Internet Things Cyber-Phys. Syst., 3: 121-154, 2023.

[4] A. Tamkin, M. Brundage, J. Clark, D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," arXiv preprint arXiv:2102.02503, 2021.

[5] Y. Yibin, C. Na, X. Chaoqian, Y. Junjian, "Accuracy assessment and analysis for GPT2," Acta Geod. Cartographica Sin., 44(7):726, 2015.

[6] A. S. George, A. H. George, "A review of ChatGPT AI's impact on several business sectors," Partners Univers. Int. Innovation J., 1(1): 9-23, 2023.

[7] X. Zheng, C. Zhang, P. C. Woodland, "Adapting GPT, GPT-2 and BERT language models for speech recognition," in Proc. 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU): 162-168, 2021.

[8] Z. Liu, R. A. Roberts, M. Lal-Nag, X. Chen, R. Huang, W. Tong, "AI-based language models powering drug discovery and development," Drug Discovery Today, 26(11): 2593-2607, 2021.

[9] N. De Cao, T. Kipf, "MolGAN: An implicit generative model for small molecular graphs," arXiv preprint arXiv:1805.11973, 2018.

[10] M. Mars, "From word embeddings to pre-trained language models: A state-of-the-art walkthrough," Appl. Sci., 12(17): 8805, 2022.

[11] Y. Qu, P. Liu, W. Song, L. Liu, M. Cheng, "A text generation and prediction system: pre-training on new corpora using BERT and GPT-2," in Proc. 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC): 323-326, 2020.

[12] O. Shatalov, N. Ryabova, "Towards Russian text generation problem using OpenAI's GPT-2," in Proc. CEUR Workshop 2021.

[13] N. Alexandr, O. Irina, K. Tatyana, K. Inessa, P. Arina, "Fine-tuning gpt-3 for russian text summarization," in Data Science and Intelligent Systems: Proceedings of 5th Computational Methods in Systems and Software, 2: 748-757, 2021.

[14] T. Goyal, J. J. Li, G. Durrett, "News summarization and evaluation in the era of gpt-3," arXiv preprint arXiv:2209.12356, 2022.

[15] S. Kumari, T. Pushphavati, "Question answering and text generation using BERT and GPT-2 model," in Proc. ICCMDE 2021 Computational Methods and Data Engineering: 93-110, 2022.

[16] A. H. George, A. S. Hameed, A. S. George, T. Baskar, "Study on quantitative understanding and knowledge of farmers in trichy district," Partners Univer. Int. Res. J., 1(2): 5-8, 2022.

[17] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[18] A. Kraft, "Triggering models: Measuring and mitigating bias in german language generation," Universität Hamburg, 2021.

[19] M. Moila, T. Modipa, J. Manamela, "The analysis of a GPT-based Sepedi text generation model," in Proc. International Conference on Intelligent and Innovative Computing Applications: 144-152, 2022.

[20] S. P. Ramalepe, T. I. Modipa, M. H. Davel, "The development of a sepedi text generation model using transformers," in Proc. Southern Africa Telecommunication Networks and Applications Conference (SATNAC)," 2022.

[21] L. Butgereit, A. van Staden, "Supporting home-language education in africa with multi-lingual mathematics tutoring using GPT-4," in Proc. International Conference on Artificial Intelligence and its Applications: 44-49, 2023.

[22] S. Yang, D. Feng, L. Qiao, Z. Kan, D. Li, "Exploring pre-trained language models for event extraction and generation," in Proc. the 57th Annual Meeting of the Association for Computational Linguistics: 5284-5294, 2019.

[23] J. Mellon, J. Bailey, R. Scott, J. Breckwoldt, M. Miori, " Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale," SSRN Electron. J., 2022.

[24] A. Radford et al., "Better language models and their implications. OpenAI," ed, 2019.

[25] R. Dale, "GPT-3: What's it good for?," Nat. Lang. Eng., 27(1): 113-118, 2021.

[26] M. Zhang, J. Li, "A commentary of GPT-3 in MIT Technology Review 2021," Fundam. Res., 1(6): 831-833, 2021.

[27] B. Zhu, Z. Gu, Y. Qian, F. Lau, Z. Tian, "Leveraging transferability and improved beam search in textual adversarial attacks," Neurocomputing, 500: 135-142, 2022.

[28] D. Kim, J. Lee, "Designing an algorithm-driven text generation system for personalized and interactive news reading," Int. J. Human–Computer Interact., 35(2): 109-122, 2019.

[29] X. He et al., "Cater: Intellectual property protection on text generation apis via conditional watermarks," Adv. Neural Inf. Process. Syst., 35: 5431-5445, 2022.

[30] J. Li, Z. Li, L. Mou, X. Jiang, M. Lyu, I. King, "Unsupervised text generation by learning from search," Adv. Neural Inf. Process. Sys., 33: 10820-10831, 2020.

Biographies

Mohadese Solouki was born 1999 in Tehran, Iran. She received her B.S. degree in Software Engineering in2020. She is currently a M.Sc. student in artificial intelligence at Islamic Azad University (Science and Research Branch) and a research fellow at the Image Processing and Natural Language Processing. Her research interests are machine learning, Deep Learning.

- Email: m.soluki1999@gmail.com
- ORCID: 0009-0008-6502-945X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Zahra Askarinejad is a faculty member of the Computer Department at the Refah College University. She holds a Master's degree in Software Engineering from APU University and a PhD in computer science in field of Software Engineering from the University Putra Malaysia. Her areas of expertise include software engineering, requirement engineering, and HCI.

- Email: askarinejad@refah.ac.ir
- ORCID: 0000-0002-7204-1384
- Web of Science Researcher ID: JDM-5453-2023
- Scopus Author ID: NA
- Homepage: https://refah.ac.ir/cv

Nastaran Zanjani is a faculty member of the Computer Department at the Refah College University. She holds a Bachelor's degree in Electrical Engineering with a specialization in Electronics from the Shahid Beheshti University. She also has a Master's degree in Telecommunications Engineering from Khaje Nasir Toosi University and a Ph.D. in Information Technology from the Queensland University of Technology in Australia. Her areas of expertise include information technology, and HCI.

- Email: m.soluki1999@gmail.com
- ORCID: 0000-0002-5307-683X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://refah.ac.ir/cv

How to cite this paper:
M. Soluki, Z. Askarinejadamiri, N. Zanjani, “Using GPT-2 model and Hazm library for persian text generation,” J. Electr. Comput. Eng. Innovations, 12(2): 285-303, 2024.

DOI: 10.22061/jecei.2023.10101.690

URL: https://jecei.sru.ac.ir/article_2023.html