JUN IA: Rewording, Grammatical Error Correction and Plagiarism detection

BIG DATA & MARKETING QUANTITATIF

Nelly AGOSSOU, Jean-Baptiste GOMEZ & Ulrich SEGODO

Student in Master 2 EBDS and MAG3; Aix-Marseille School of Economics (AMSE)

Final Version

January 2024

1. Introduction

Artificial intelligence (AI), a branch of computer science that aims to create machines that can reason and learn like humans. AI has come a long way in recent years, and it is now used in many fields, including research, medicine, finance, and industry. One of the most promising applications of AI is natural language processing (NLP). NLP is the study of the interactions between computers and natural languages. One of the challenges of NLP is the processing of textual data. They can be documents, emails, tweets, online comments, etc. Textual data processing is necessary for many applications, such as machine translation, speech recognition, and information retrieval. Textual data has evolved significantly in recent years. The amount of textual data available has exploded, and the complexity of that data has also increased. This development has posed new challenges for the processing of textual data. Another challenge is the need to process unstructured textual data. Most textual data is unstructured, meaning it's not organized in an orderly fashion. Unstructured textual data is difficult to process because it is difficult to understand its meaning.

The JUN AI project aims to address these challenges. JUN AI is a GPT-3.5-based solution for text generation. GPT-3.5 is a pre-trained language model developed by OpenAI. It can generate high-quality text, and it is available in multiple languages.

JUN AI distinguishes itself from other products on the market in the following ways:

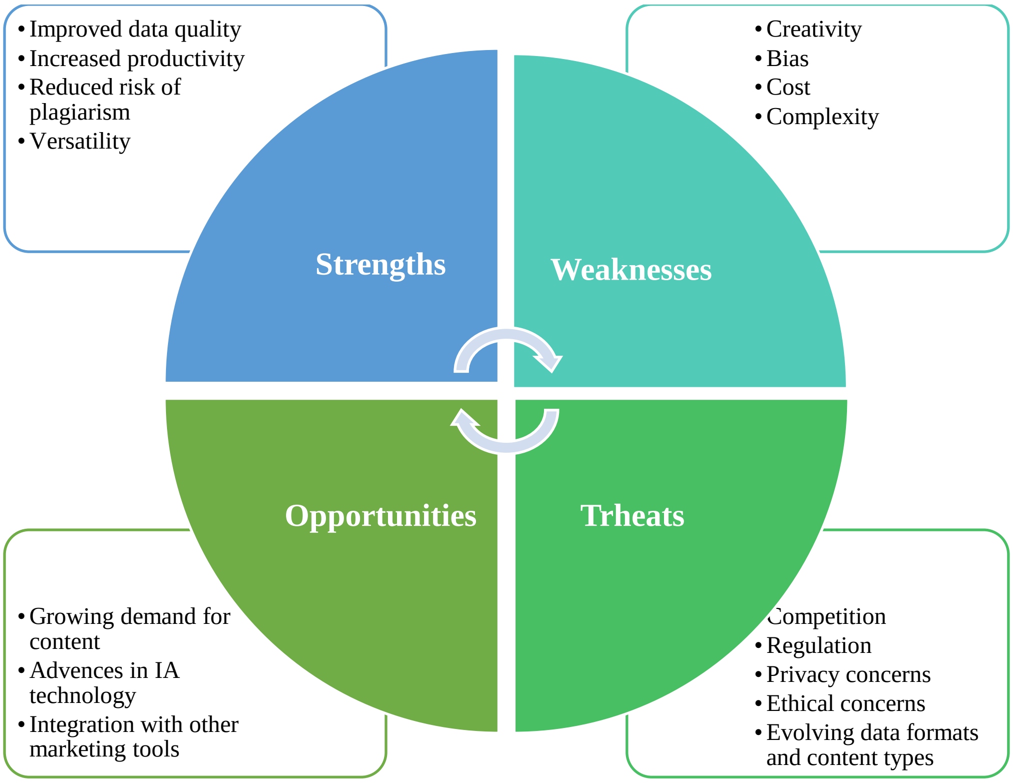
- It is cross-platform and accessible from a phone. It is also available in multiple languages.
- It is accessible to everyone at an affordable cost, starting from a daily, weekly, monthly, or annual subscription.
- It is good value for money, as it saves time and increases user productivity.
- It offers several features such as a dialog box, plagiarism detection, and more.

In the construction of this project, we intend to start with the grammatical checking and error correction part, then the rephrasing and then compare its results with that of the Bert algorithm.

Grammatical checking and error correction is the task of detecting and correcting grammatical errors in a text. This task is important because it helps to ensure the quality of the text. Rephrasing is the task of rephrasing a text without changing its meaning. This task is useful for making the text clearer, more concise, or more engaging. By comparing the results of JUN AI with those of Bert, we will be able to evaluate the performance of JUN AI for the tasks of grammatical checking, error correction and rephrasing.

2. Marketing solution

2.1. SWOT analysis



- Improved data quality
- Increased productivity
- Reduced risk of plagiarism
- Versatility

- Creativity
- Bias
- Cost
- Complexity

- Growing demand for content
- Advences in IA technology
- Integration with other marketing tools

- Competition
- Regulation
- Privacy concerns
- Ethical concerns
- Evolving data formats and content types

2.2. Marketing Mix (4P)

Our diverse target audience encompasses businesses, marketing professionals, writers, content creators, and students. We are committed to delivering personalized customer service, providing tailored advice for each sector. Emphasizing customer satisfaction, we also offer dedicated technical support to address specific challenges encountered in using our services. This approach strengthens our position as a reliable partner, addressing the varied needs of our clients and underscoring our commitment to the success of each user, whether in the professional, marketing, creative, or educational realms.

Product:

- Rewords and paraphrases text in a natural and engaging way
- Improves the clarity and conciseness of text
- Generates new and original content
- Corrects grammatical errors

Price:

- Premium pricing model: Offer a free basic plan with limited features and charge a premium for a plan with more features. Notice that there is a subscription for individuals and professionals.
- Enterprise pricing model: Offer custom pricing for large businesses that need a high volume of rewording.

Place:

Distribution channels:

- Direct sales: Sell JUN AI directly to businesses through a website or online store.
- Partnerships: Partner with other businesses that offer complementary products or services.
- Online marketplaces: Sell Rewording AI on online marketplaces like Google Marketplace, Microsoft Azure Marketplace, AWS Marketplace, and IBM Marketplace.

Promotion:

Marketing channels:

- Content marketing: Create and publish blog posts, articles, and infographics that educate potential customers about the benefits of Rewording AI.
- Social media marketing: Use social media platforms like LinkedIn, Instagram, Twitter, Facebook and Tiktok to connect with potential customers and promote JUN AI. ...), Organize a contest or giveaway.
- Paid advertising: Use paid advertising platforms like Google Ads and LinkedIn Ads to reach a wider audience.
- Public relations: Generate positive press coverage for JUN AI by pitching stories to journalists and bloggers.

3. Data Exploration

Our database comes from Kaggle. This is the JFLEG Dataset: An English Grammatical Error Benchmark. The JFLEG (JHU FLuency-Extended GUG) dataset is a comprehensive benchmark for English Grammatical Error Correction (GEC) systems. It serves as a gold standard for developing and evaluating the effectiveness of GEC systems in terms of fluency and grammaticality in English texts. The dataset is specifically designed to assess the native-sounding quality and grammatical precision of written English sentences. This database contains 2 files namely: validation.csv (755 observations) and test.csv (748 observations). The first file is used to evaluate the performance of our grammatical error correction model on different text types. This allows you to analyze how accurately our model corrects grammatical errors compared to its ground truth or baseline corrections. provided in this file. The second file is used to perform tests on unseen data or to validate the generalization capabilities of our trained model. It provides another set of sentence correction pairs that allow comparison with established reference standards. The structure of these two files is as follows:

Column name	Description
Sentence	This column contains the original English sentences that may contain grammatical errors. (Text)
Corrections	This column contains the corrected versions of the sentences in the sentence column, where the grammatical errors have been fixed. (Text)

4. Result

In the context of Grammar Correction, we employ the T5 Model exclusively designed for the English language. Developed by Google AI, T5 is available for public download and use. This encoder-decoder model transforms all Natural Language Processing (NLP) problems into a text-to-text format, utilizing a training approach known as teacher forcing. This entails always having an input sequence along with its corresponding target sequence during the training process.

For the implementation, we will utilize the Python package named Happy Transformer. Happy Transformer is constructed on top Hugging Face's Transformers library, simplifying the implementation and training of transformer models with just a few lines of code.

We'll evaluate the model before and after fine-tuning using a common metric called loss. Loss can be described as how "wrong" the model's predictions are compared to the correct answers. So, if the loss decreases after fine-tuning, then that suggests the model learned. In our case, we obtain this result:

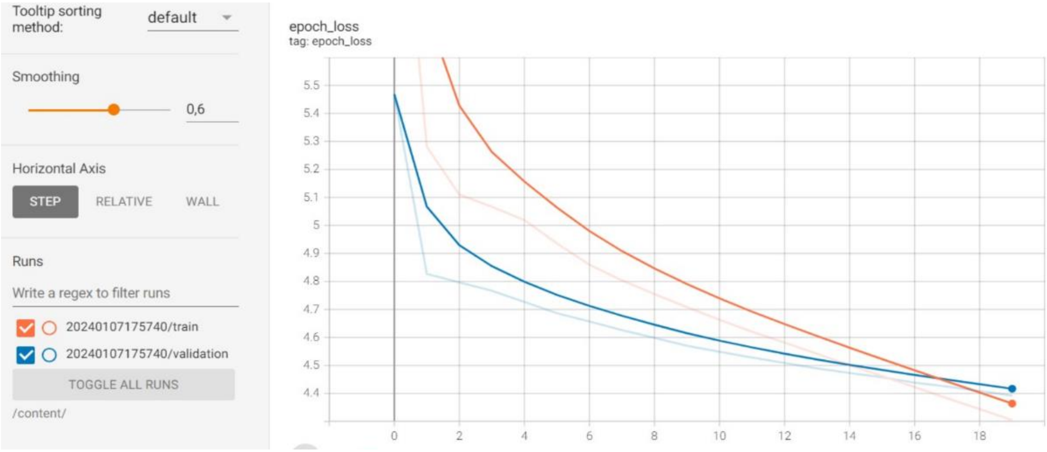
Before Loss	After Loss
1.28	0.47

As we can see the loss decreased so the model learned.

Epoch	Training Loss	Validation Loss
1	0.619800	0.622075
34	0.405600	0.615530
68	0.603100	0.564905
102	0.665900	0.534906
136	0.594300	0.523595
170	0.622100	0.505431
204	0.565100	0.504468
238	0.596200	0.491850
272	0.580400	0.491327
306	0.657300	0.487521
340	0.560500	0.487402

Analyzing the respective evolution of the cost function for both the training data and the validation data reveals that, starting from **epoch 68**, the cost functions of both datasets consistently decrease until **epoch 340**. This consistent decrease is crucial in the context of minimizing the cost function for our model. However, it's important to note that the model is sensitive to overfitting.

Within the notebook, we explore additional deep learning algorithms such as LSTM (Long Short-Term Memory) and RNN (Recurrent Neural Network) models. In the following sections, we will provide a brief overview of the results obtained from the LSTM model. As we can see for the LSTM model, we don't have overfitting problem, the cost functions of both datasets consistently decrease. Same for the evaluation loss with each iteration base on validation dataset.



Tooltip sorting method: default

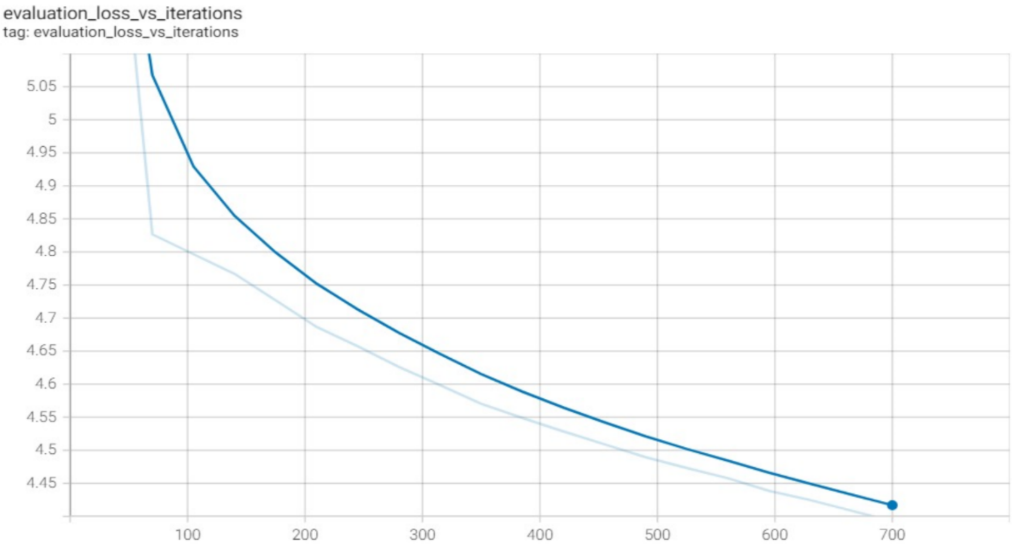
Smoothing 0,6

Horizontal Axis: STEP

☑ ○ 20240107175740/train

☑ ○ 20240107175740/validation

/content/



However, for the evaluation of this model we use the **GLUE** score **(General Language Understanding Evaluation).** The GLUE score of train data is **3.15%** and for the validation data **2.96%** is still very weak. We can conclude that the LSTM model is not a high-performance model for our dataset.

Inference:

Now, let's leverage the trained model to correct the grammar of examples provided to it. For this task, we will employ the **`generate_text()`** method from the **`happy_tt`** library. Additionally, we will utilize an algorithm known as beam search for the text generation process. We took 2 incorrect sentences, passed them to our model for correction and compared with **GPT 3.5 result.**

Test 1:

Incorrect sentence: "grammar: This sentences, has bads grammar and spelling!"

HTT correct sentence: “This sentences, has bad grammar and spelling!”

GPT-3.5 correct sentence: "This sentence has bad grammar and spelling."

Test 2:

Incorrect sentence: "grammar: I am enjoys, writtings articles ons AI."

HTT correct sentence: “I am enjoying writing articles on AI.”

GPT-3.5 correct sentence: "I enjoy writing articles on AI."

The correct sentence provides by the T5 model and Openai (GPT 3.5) model is not the same, but both is correct. Also we have explored more advanced transformer-based models. However, constraints in hardware have hindered our ability to train them with our existing database. These limitations stem not only from the limited computational power of our computers but also from the heightened complexity of these advanced models. The computational resource requirements and sophistication of these new models surpass the capabilities of our current infrastructure. We are actively exploring solutions to overcome these challenges, aiming to fully leverage the potential of these advanced models in our future work.

5. Conclusion

In conclusion, Jun AI (Rewording, Grammatical Error Correction, and Plagiarism Detection) is a project poised for long-term performance. Given the time at our disposal, we have successfully completed the grammatical error correction aspect, which is one of the fundamental pillars of this project. Following the implementation of various models (RNN, LSTM, T5 Model), we can highlight the transformer-based T5 model (Happy Transformer) and the OpenAI model based on GPT 3.5 for our product. This achievement already addresses three strengths of our product, namely improved data quality (the enhancement of the overall quality of the data used in the Jun AI project. In the context of grammatical error correction and plagiarism detection, improving data quality could involve refining the training data used for machine learning models.), versatility (the system can effectively address different types of grammatical errors, diverse writing styles, and a range of content. For the moment only English sentences), and increased productivity for the customer (contributes to saving time and effort for its users. In the context of grammatical error correction and plagiarism detection, a highly productive tool like our product, would efficiently and accurately identify and correct errors, thereby reducing the time and manual effort required for proofreading).