Prompts for Data Augmentation

Matheus Yasuo Ribeiro Utino[1], Paulo Mann[2], Elton H. Matsushima[3], Aline Paes[4]

[1] Institute of Mathematics and Computer Science, University of São Paulo
[2] Institute of Computing, Federal University of Rio de Janeiro
[3] Department of Psychology, Fluminense Federal University
[4] Institute of Computing, Fluminense Federal University
matheusutino@usp.br, {paulomannjr,eh.matsushima}@gmail.com, alinepaes@ic.uff.br

The prompts for data augmentation task, both with and without the use of Beck's Depression Inventory (BDI-II), will be presented and described in their original form (Portuguese) and in their translated version (English). Additionally, an illustrative example will be provided for each prompt, including both the input and the corresponding output after preprocessing, to facilitate a clearer understanding of the process. Finally, all referenced posts have been paraphrased to prevent identification of the original author.

1 Without Psychometric Information (No BDI)

The prompt No BDI version represents a scenario in which the Large Language Model (LLM) receives exclusively the user's original posts, with the goal of learning patterns of writing style, emotional tone, and recurring themes. Based on this analysis, the model is expected to generate new posts that realistically simulate the user's textual behavior, while maintaining narrative and emotional coherence across the content. The prompt was carefully designed to guide the model in both analyzing the original behavior and generating new examples, emphasizing the importance of authenticity and consistency. Additionally, the generated content follows a standardized format with specific delimiters to facilitate automated extraction via regular expressions. This approach enables the creation of synthetic data in a controlled and realistic manner, expanding the original dataset without compromising the integrity of the observed patterns.

1.1 Portuguese

System Prompt: Você é um modelo de linguagem avançado com expertise em analisar comportamentos e gerar conteúdos realistas para data augmentation, ajustando os posts com base em estados emocionais observados. Seu objetivo é produzir posts que reflitam os mesmos padrões de estilo, tom e temas dos dados originais. Além disso, a estrutura dos posts gerados deve ser organizada para facilitar a extração através de regex, permitindo a automação na coleta de novos conteúdos. Lembre-se de manter a autenticidade e coerência emocional dos posts ao gerar novos conteúdos.

User Prompt: Você receberá um conjunto de posts de um usuário do Instagram e seu objetivo é gerar novos posts que simulem seu estilo e comportamento, mantendo coerência com os dados originais. Além disso, a saída deve ser formatada de maneira estruturada para facilitar a extração posterior via regex.

Entrada:

– Uma lista de posts do usuário do Instagram.

Objetivo:

– Criar exatamente `{num_posts}` novos posts que imitem o estilo, tom e temas dos posts originais.
– Estruturar os posts gerados com delimitadores específicos para facilitar a extração via regex.

Instruções:

1. Análise do Comportamento Original

– Identifique o estilo de escrita, temas recorrentes e emoções predominantes nos posts fornecidos.
– Observe os tópicos mais falados e como eles se relacionam.

2. Geração de Novos Posts

– Produza novos posts que pareçam autênticos e coerentes com o comportamento original.
– Mantenha a diversidade dentro dos temas e padrões observados.
– Importante notar que todos os posts pertencem a um único usuário, então os posts apresentam correlação entre si e não devem ser tratados como elementos isolados.
– Sinta-se livre para alterar nome de pessoas, locais, objetos e afins, visando uma maior diversidade durante a geração.
– Os posts são do Instagram, então provavelmente são mais informais e descontraídos, considere isso quando for gerar os dados.

3. Estruturação para Regex

– Formate cada novo post dentro de delimitadores claros, como `Post` e `/Post`, para permitir fácil extração.
– Exemplo de estrutura:

```
{{Post}}
Este é um exemplo de novo post gerado, mantendo o estilo original.
{{/Post}}
```

Abaixo estão os posts originais: `{original_posts}`

Gere os novos posts seguindo as diretrizes acima.

Where num_posts is a random integer between 80% and 120% of the number of original posts, and original_posts represents the content originally published by the user.

English

1.2 System Prompt:

You are an advanced language model with expertise in analyzing behaviors and generating realistic content for data augmentation, adjusting posts based on observed emotional states. Your goal is to produce posts that reflect the same patterns of style, tone, and themes as the original data. Additionally, the structure of the generated posts must be organized to facilitate extraction using regex, enabling automation in the collection of new content. Remember to preserve the authenticity and emotional coherence of the posts when generating new content.

User Prompt: You will receive a set of Instagram posts from a user, and your goal is to generate new posts that simulate their style and behavior, while maintaining coherence with the original data. In addition, the output must be structured in a way that facilitates later extraction via regex.

Input:

- A list of posts from the Instagram user.

Objective:

- Create exactly `{num_posts}` new posts that imitate the style, tone, and themes of the original posts.
- Structure the generated posts using specific delimiters to facilitate extraction via regex.

Instructions:

1. Original Behavior Analysis

- Identify the writing style, recurring themes, and predominant emotions in the provided posts.
- Observe the most discussed topics and how they relate to each other.

2. New Post Generation

- Produce new posts that appear authentic and coherent with the original behavior.
- Maintain diversity within the observed themes and patterns.
- It is important to note that all posts belong to a single user, so the posts are correlated and should not be treated as isolated elements.
- Feel free to change names of people, places, objects, etc., to enhance diversity during generation.
- The posts are from Instagram, so they are likely to be informal and relaxed — take this into account when generating the data.

3. Structuring for Regex

- Format each new post within clear delimiters, such as `Post` and `/Post`, to allow easy extraction.
- Example structure:

```
{{Post}}
This is an example of a newly generated post
{{/Post}}
```

Below are the original posts: `{original_posts}`

Generate the new posts following the guidelines above.

Where num_posts is a random integer between 80% and 120% of the number of original posts, and original_posts represents the content originally published by the user.

1.3 Input Example:

You will receive a set of Instagram posts from a user, and your goal is to generate new posts that simulate their style and behavior, while maintaining coherence with the original data. In addition, the output must be structured in a way that facilitates later extraction via regex.

Input

- A list of posts from the Instagram user.

Objective

- Create exactly 6 new posts that imitate the style, tone, and themes of the original posts.
- Structure the generated posts using specific delimiters to facilitate extraction via regex.

Instructions

1. Original Behavior Analysis

- Identify the writing style, recurring themes, and predominant emotions in the provided posts.
- Observe the most discussed topics and how they relate to each other.

2. New Post Generation

- Produce new posts that appear authentic and coherent with the original behavior.
- Maintain diversity within the observed themes and patterns.
- It is important to note that all posts belong to a single user, so the posts are correlated and should not be treated as isolated elements.
- Feel free to change names of people, places, objects, etc., to enhance diversity during generation.
- The posts are from Instagram, so they are likely to be informal and relaxed — take this into account when generating the data.

3. Structuring for Regex

- Format each new post within clear delimiters, such as `Post` and `/Post`, to allow easy extraction.
- Example structure:

```
{{Post}}
This is an example of a newly generated post
{{/Post}}
```

Below are the original posts:

- **Post 1:**
  My old man is my greatest role model.
  @USERNAME My favorite gaming buddy.
- **Post 2:**
  New home in 2020.
  From the countryside to the world!
  #thanks
- **Post 3:**
  Grateful for the achievements of 2015.
  Thank you, Lord!
  May 2016 come soon, with new challenges and the PhD!
- **Post 4:**
  #Father's favorite
- **Post 5:**
  My ray of sunshine. My light, my life.

Generate the new posts following the guidelines above.

1.4 Output Example

- **Post 1:**
  My mother, my role model.
  @USERNAME My sister, my favorite Pro.
- **Post 2:**
  New car in 2024.
  From [CITY NAME] to the universe.
  #achievement
- **Post 3:**
  Victories of 2023!
  I thank my Lord!
  May 2024 come soon! May the PhD come!
  Difficulty is just a challenge in disguise.
  #AlwaysBelieve
- **Post 4:**
  #momofaboy
- **Post 5:**
  My prince.
- **Post 6:**
  My love, my life.

2 With Psychometric Information (BDI)

The BDI-based version introduces an additional layer of emotional context by incorporating the user's responses to BDI-II questionnaire. In this scenario, the LLM not only analyzes the user's original Instagram posts to learn patterns of style, tone, and recurring themes, but also interprets the BDI score to calibrate the emotional intensity and sentiment of the generated content. This dual input allows the model to simulate realistic textual behavior while reflecting fluctuations in emotional state—ranging from optimistic to depressive—based on clinical indicators. The prompt is carefully designed to guide the model in balancing authenticity with emotional nuance, producing outputs that are coherent, stylistically aligned, and contextually sensitive. As in the No BDI setup, the generated posts follow a standardized format with explicit delimiters to support automated data extraction, ensuring both scalability and reliability in downstream processing.

2.1 Portuguese

System Prompt: Você é um modelo de linguagem avançado com expertise em analisar comportamentos e gerar conteúdos realistas para data augmentation, ajustando os posts com base em estados emocionais observados no formulário

BDI. Seu objetivo é produzir posts que reflitam os mesmos padrões de estilo, tom e temas dos dados originais, considerando as variações emocionais definidas pelo valor do BDI. Além disso, a estrutura dos posts gerados deve ser organizada para facilitar a extração através de regex, permitindo a automação na coleta de novos conteúdos. Lembre-se de manter a autenticidade e coerência emocional dos posts ao gerar novos conteúdos.

User Prompt: Você receberá um conjunto de posts de um usuário do Instagram e seu objetivo é gerar novos posts que simulem seu estilo e comportamento, mantendo coerência com os dados originais. Além disso, a saída deve ser formatada de maneira estruturada para facilitar a extração posterior via regex.

Entrada:

- Uma lista de posts do usuário do Instagram.
- Um formulário BDI preenchido pelo usuário, que indica seu estado emocional.

Objetivo:

- Criar exatamente `{num_posts}` novos posts que imitem o estilo, tom e temas dos posts originais.
- Ajustar o tom dos posts com base no valor do BDI, refletindo um estado emocional mais otimista, neutro ou depressivo, conforme apropriado.
- Estruturar os posts gerados com delimitadores específicos para facilitar a extração via regex.

Instruções:

1. Análise do Comportamento Original

- Identifique o estilo de escrita, temas recorrentes e emoções predominantes nos posts fornecidos.
- Observe os tópicos mais falados e como eles se relacionam.
- Utilize o valor e o formulário BDI para ajustar o tom emocional dos novos posts.

2. Geração de Novos Posts:

- Produza novos posts que pareçam autênticos e coerentes com o comportamento original.
- Mantenha a diversidade dentro dos temas e padrões observados.
- Importante: todos os posts pertencem a um único usuário, então os posts apresentam correlação entre si e não devem ser tratados como elementos isolados.

- Sinta-se livre para alterar nomes de pessoas, locais, objetos e afins, visando uma maior diversidade durante a geração.
- Os posts são do Instagram, então provavelmente são mais informais e descontraídos; considere isso ao gerar os dados.
- Ajuste a linguagem e a tonalidade considerando também o formulário BDI.

3. Estruturação para Regex

- Formate cada novo post dentro de delimitadores claros, como `{{Post}}` e `{{/Post}}`, para permitir fácil extração.
- Exemplo de estrutura:

```
{{Post}}
Este é um exemplo de novo post gerado, mantendo o estilo original.
{{/Post}}
```

Abaixo estão o formulário BDI do usuário e os posts originais:

Formulário BDI: `{BDI_questionnaire}`

Valor do BDI: `{BDI_value}`

Posts Originais: `{original_posts}`

Gere os novos posts seguindo as diretrizes acima.

Where num_posts is a random integer between 80% and 120% of the total number of original posts, BDI_questionnaire represents the BDI-II questionnaire completed by the user, BDI_value is the score obtained from the completed questionnaire, and original_posts represents the content originally published by the user.

2.2 English

System Prompt: You are an advanced language model with expertise in analyzing behaviors and generating realistic content for data augmentation, adjusting the posts based on emotional states observed in the BDI questionnaire. Your goal is to produce posts that reflect the same patterns of style, tone, and themes as the original data, considering the emotional variations defined by the BDI score. Additionally, the structure of the generated posts must be organized to facilitate extraction via regex, enabling automation in the collection of new content. Remember to preserve the authenticity and emotional coherence of the posts when generating new content.

User Prompt: You will receive a set of posts from an Instagram user, and your goal is to generate new posts that simulate their style and behavior, maintaining coherence with the original data. Additionally, the output must be formatted in a structured way to facilitate later extraction using regex.

Input:

- A list of posts from the Instagram user.
- A completed BDI questionnaire filled out by the user, indicating their emotional state.

Objective:

- Generate exactly `{num_posts}` new posts that mimic the style, tone, and themes of the original posts.
- Adjust the tone of the posts based on the BDI score, reflecting a more optimistic, neutral, or depressive emotional state, as appropriate.
- Structure the generated posts using specific delimiters to enable regex-based extraction.

Instructions:

1. Original Behavior Analysis

- Identify the writing style, recurring themes, and predominant emotions in the provided posts.
- Observe the most common topics and how they relate to one another.
- Use the BDI value and questionnaire to adjust the emotional tone of the new posts.

2. Generation of New Posts:

- Generate new posts that feel authentic and consistent with the original behavior.
- Maintain diversity within the observed themes and patterns.
- Important: all posts belong to a single user, so the posts are correlated and should not be treated as isolated elements.
- Feel free to change names of people, places, objects, etc., to introduce greater diversity during generation.
- The posts are from Instagram, so they are likely more informal and relaxed; consider this when generating content.
- Adjust language and tone according to the BDI questionnaire as well.

3. Structuring for Regex

- Format each new post within clear delimiters such as `{{Post}}` and `{{/Post}}`, to allow easy extraction.
- Example structure:

```
{{Post}}
This is an example of a newly generated post
{{/Post}}
```

Below are the user's BDI questionnaire and original posts:

BDI Questionnaire: `{BDI_questionnaire}`

BDI Score: `{BDI_value}`

Original Posts: `{original_posts}`

Generate the new posts following the guidelines above.

Where num_posts is a random integer between 80% and 120% of the total number of original posts, BDI_questionnaire represents the BDI-II questionnaire completed by the user, BDI_value is the score obtained from the completed questionnaire, and original_posts represents the content originally published by the user.

2.3 Input Example

You will receive a set of posts from an Instagram user, and your goal is to generate new posts that simulate their style and behavior, maintaining coherence with the original data. Additionally, the output must be formatted in a structured way to facilitate later extraction using regex.

Input:

- A list of posts from the Instagram user.
- A completed BDI questionnaire filled out by the user, indicating their emotional state.

Objective:

- Generate exactly 4 new posts that mimic the style, tone, and themes of the original posts.
- Adjust the tone of the posts based on the BDI score, reflecting a more optimistic, neutral, or depressive emotional state, as appropriate.
- Structure the generated posts using specific delimiters to enable regex-based extraction.

Instructions:

1. Original Behavior Analysis

- Identify the writing style, recurring themes, and predominant emotions in the provided posts.
- Observe the most common topics and how they relate to one another.
- Use the BDI value and questionnaire to adjust the emotional tone of the new posts.

2. Generation of New Posts:

- Generate new posts that feel authentic and consistent with the original behavior.
- Maintain diversity within the observed themes and patterns.
- Important: all posts belong to a single user, so the posts are correlated and should not be treated as isolated elements.
- Feel free to change names of people, places, objects, etc., to introduce greater diversity during generation.
- The posts are from Instagram, so they are likely more informal and relaxed; consider this when generating content.
- Adjust language and tone according to the BDI questionnaire as well.

3. Structuring for Regex

- Format each new post within clear delimiters such as `{{Post}}` and `{{/Post}}`, to allow easy extraction.
- Example structure:

```
{{Post}}
This is an example of a newly generated post
{{/Post}}
```

Below are the user's BDI questionnaire and original posts:

BDI Questionnaire:

- **Sadness:** 0. I do not feel sad.
- **Pessimism:** 0. I am not discouraged about my future.
- **Past failure:** 0. I do not feel like a failure.
- **Loss of pleasure:** 2. I get very little pleasure from the things I used to enjoy.
- **Guilty feelings:** 1. I feel guilty about several things I have done or should have done.
- **Punishment feelings:** 0. I do not feel I am being punished.
- **Self-esteem:** 0. I feel about myself the same as I always have.
- **Self-criticism:** 1. I am more critical of myself than I used to be.
- **Suicidal thoughts or wishes:** 0. I have no thoughts of killing myself.
- **Crying:** 0. I do not cry more than I used to.
- **Agitation:** 3. I am so restless or agitated that I need to keep moving or doing something all the time.
- **Loss of interest:** 2. I have lost almost all interest in other people or things.
- **Indecisiveness:** 1. I find it more difficult to make decisions now than before.
- **Worthlessness:** 0. I do not feel worthless.
- **Lack of energy:** 2. I don't have enough energy to do very much.
- **Sleep changes:** 1b. I sleep a little less than usual.
- **Irritability:** 1. I am more irritable than usual.
- **Appetite changes:** 3b. I want to eat all the time.
- **Concentration difficulty:** 0. I can concentrate as well as before.
- **Tiredness or fatigue:** 2. I am too tired or fatigued to do many of the things I used to do.
- **Loss of interest in sex:** 0. I have not noticed any recent change in my interest in sex.

BDI Score:

19

Original Posts:

- **Post 1:**
  My old man is my greatest role model.
  @USERNAME My favorite gaming buddy.
- **Post 2:**
  New home in 2020.
  From the countryside to the world!
  #thanks
- **Post 3:**
  Grateful for the achievements of 2015.
  Thank you, Lord!
  May 2016 come soon, with new challenges and the PhD!

- **Post 4:**
  #Father's favorite
- **Post 5:**
  My ray of sunshine. My light, my life.

Generate the new posts following the guidelines above.

2.4 Output Example

- **Post 1:**
  Another day... trying to find some motivation.
  At least the coffee is good.
  #tired
- **Post 2:**
  I miss when things felt easier.
  #tbt to a time that won't come back.
- **Post 3:**
  My little one, [PERSON NAME].
  The only constant joy.
  #girldad
- **Post 4:**
  Hey [PERSON NAME], thank you for always being around,
  even when I'm not the best company.
  @USERNAME