Emotion-Aware Language Models: A Self-Assessment Study

Viktor Due Pedersen
vipe@itu.dk

15 December 2023, Copenhagen, Denmark

Abstract

Obtaining or annotating data can for complex and low-resource domains be a difficult task. One solution to this problem is to augment previous annotated texts using large language models (LLMs). This paper investigates the capabilities of LLMs, specifically GPT-4 and Llama2 70b, in the context of data augmentation for texts expressing various social dimensions. My study focuses on three key areas: semantic and lexical differences in augmentations, LLMs' ability to classify social dimensions in augmented texts, and the agreement between models on the ranking of these texts based on how the text conveys a social dimension. The results reveal that both GPT-4 and Llama2 effectively augment texts with high semantic similarity and lexical diversity. GPT-4 demonstrates a stronger ability to classify augmented data, whereas Llama2 tends to produce more diverse augmentations, perhaps limiting the possibility to distinguish the texts. The two models have a high agreement in ranking the quality of augmentations, though discrepancies exist in interpreting certain dimensions. These insights contribute to understanding the subtleties and effectiveness of LLMs in text augmentation and emotion expression, highlighting potential applications and limitations in data augmentation for augmentation tasks with subtle label differences.

1 Introduction & Related work

With the rise of large language models (LLMs), a vast number of use-cases have arisen. An important field is that of data augmentation for data where obtaining or annotating it can be difficult. Augmentation for texts has long been a difficult topic in which simple methods like character modification [1], deliberately adding misspellings [2] and substituting synonyms [3] have been applied. More advanced methods that consider context have been applied, such as adding <mask> tokens at random places and replacing or inserting tokens with BERT [4, 5]. Efforts involving back-and-forth translation have also been tested [6]. A common drawback of these methods is that they simply modify parts of the text, adding noise, without expanding or rephrasing the content.

Autoregressive transformers like GPT-4 [7] and Llama2 [8] allow the use of off-the-shelf models. These can be provided with context to generate entirely new texts. Several studies have examined LLMs ability to augment data [5, 9–11], and tested the methods on downstream classification tasks. One aspect is the ability to augment data. However, when it comes to more subtle nuances of text generation, such as creating text for specific social dimensions, our current approaches needs refinement. We must examine the texts themselves to assess whether the LLM understands the differences between the labels. The SOCKET Benchmark [12] was introduced exactly for the purpose of assesing LLMs ability to understand social language.

The authors of [11] conducted an analysis on classification models trained on varying amounts of human-labeled and synthetically augmented data. They found that human-annotated data generally outperform augmented data for binary balanced tasks, sensitive tasks and multiclass balanced tasks. They found high performance on unbalanced multi-

class tasks, but used a sample strategy where they oversampled minority classes. Additionally they found that augmentations conducted by Llama2 70b had a greater diversity than augmentations generated by GPT-4, and argue that this can prove beneficial for multiclass unbalanced tasks.

This paper aims to independently evaluate the data augmentations from [11], separate from any downstream task. It examines whether Llama2 and GPT-4 can assess the quality of their own augmentations and if the models agree on the labeling of texts with subtle distinctions.

My contributions overall are:

Section 4.1 Explore the semantic and lexical differences between augmentations performed by GPT-4 and Llama. I demonstrates that GPT-4 and Llama can both synthetically augment texts, maintaining high semantic similarity and lexical diversity compared to the original texts.

Section 4.2 Explores GPT-4 and Llama's ability to distinguish each of the social dimensions described in table 1. I demonstrate that both LLMs can distinguish subtle dimensions in augmentations, even those created by the other model. Llama generally finds it easier to distinguish augmentations made by GPT-4 than those made by itself.

Section 4.3 Explores whether GPT-4 and Llama agree on how well an augmented text express a given dimension from section table 1. I find that both LLMs, at minimum, rank the quality of augmentations with no correlation, but generally agree on their rankings.

2 Data

The main dataset for this project was initially introduced in [13]. The data consists of texts gathered from social media, and annotated with one or more social dimensions. The analysis I do is not performed on the original dataset but rather a dataset introduced in [10]. They modified the social dimensions such that *similarity* and *identity* was merged, *romance* removed and *neutral* introduced. Each social dimension, from now on called label, are described in table 1.

The work introduced in [10] are continued in [11], where they, using GPT-4 and Llama2 70B, augment the texts. They construct prompts consisting of an example from the original data along with its corresponding label. The LLMs is then asked to generate 9 new texts expressing the same label. To ensure substantial differences they augment using a temperature of 1 resulting in more stochasticity. The resulting augmentations forms the basis of my experiments in section 4.2 and 4.3. Note that when the figures state *gpt-subset* the augmented data are augmented using GPT-4 and vice versa for Llama.

In section 4.1, I examine multiple datasets, all augmented with the same methodology [11]. Almost all the datasets are a part of the the SOCKET Benchmark [12]. The datasets are identification of politeness [14], presence of empathy [15], hyperbole retrieval [16], level of intimacy in online questions [17], offensive language detection in Danish [18], Talkdown: A corpus for condescension detection [19], Crowdflower sentiment analysis [20], ten social dimension [10, 11, 13], sentiment analysis [21] and finally whether two stances are at the same side of an argument [22]. The ten-dimensions dataset is used for all three experiments, where SOCKET is only used for the first.

Label	Description
Knowledge	Exchange of ideas or information.
Power	Having power over the behavior and outcomes of another.
Respect	Conferring status, appreciation, gratitude, or admiration upon another.
Trust	Will of relying on the actions or judgments of another.
Social support	Giving emotional or practical aid and companionship.
Similarity identity	Shared interests, motivations, outlooks or Shared sense of belonging to the same community or group.
Fun	Experiencing leisure, laughter, and joy.
Conflict	Contrast or diverging views.
Neutral	Neutral communication or disagreement among annotators.

Table 1: Description of the ten social dimensions

3 Methods

3.1 Llama & GPT

The chosen LLMs for this experiment are Llama and GPT. Specifically, references to Llama or GPT in this context always imply Llama2 70b and GPT-4, respectively. The Llama model used is `meta-llama/llama-2-70b-chat-hf` from `huggingface_hub`, and GPT-4 is accessed via `ChatOpenAI(model="gpt-4")` from `langchain.chat_models`.

Both GPT and Llama are Autoregressive Transformer models, built on the transformer architecture [23]. This architecture allows the models to effectively process long texts and maintain context throughout. The transformer architecture, as introduced in [23], has become a standard for large language models, especially due to its ability to handle long-range dependencies within text. This is a significant improvement over earlier architectures, such as recurrent neural networks, which, while addressing similar issues, have been outperformed by the attention mechanism.

The autoregressive nature of these models is crucial for the experiments conducted. Text generation occurs token by token, with each new token depending solely on the preceding tokens. When prompting the models, it is important to ensure high-quality, coherent output that is consistent with both one-shot inputs and the provided context.

3.2 Cosine similarity and token overlap

When examining the semantic and lexical differences in augmentations produced by Llama and GPT, I investigate two metrics, which are briefly introduced here.

Semantic similarity is measured by calculating cosine similarity, a metric that quantifies the similarity between two vectors. In this context, it is used to compare the similarity between two text embeddings: one from the original text and the other from its augmentation, both embedded using `intfloat/e5-base` [24]. Cosine similarity is defined as $\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$, where $\mathbf{A}$ and $\mathbf{B}$ represent the embeddings.

To assess lexical overlap, I calculate the fraction of tokens that are common to both texts. I define it as $\frac{|A \cap B|}{|A \cup B|}$, where A and B are the sets of tokens in the respective texts.

3.3 Prompting

With the rise of LLMs prompt-engineering has become a new field of study. With trial and error and proper guidance from `promptingguide.ai` [25], I created the prompts seen in the appendix in figure 5 and 6. These promt serves as the input to GPT and Llama in order to conduct the experiments

According to their guideline a good prompt contain the following things:

- Instruction - a specific task or instruction you want the model to perform.
- Context - external information or additional context that can steer the model to better responses.
- Input Data - the input or question that we are interested to find a response for.
- Output Indicator - the type or format of the output.

I have two different experiments, each requiring its own prompt.

In section 4.2, I conduct an experiment of an LLMs ability to distinguish texts expressing a range of labels. Initially I provide the system context. First the model is instructed that it is a classifying model and that it is going to classify based on user-inputted text. I then describe a property that all the inputted texts have, namely that it express a label. I proceed to give the model context by describing the nine different labels that can occur and that it should classify based on.

In section 4.3, the LLMs are asked to rank to what degree the texts express a given label by comparing all pairs of text and picking the text that express the label the most. An almost identical system context was provided, but through trial-and-error setup it became clear that the ranking experiment needed additional context. I stress that both inputted text express the same label, but that it has to pick a winner, i.e. draw is not an option. It became apparent that the model was not consistent in the output format so I add examples of how a good output would look like i.e. an output indicator.

For both experiments the user input is similar. I ask the specific question for this experiment, provide a label, and the texts that the question is regarding. The first experiment of distinguishing the labels, additionally require an example of how a text that express the given label could look like. Providing examples are a so-called one-shot where the model receive additional information about the task. The second experiment is zero-shot experiment as I do not provide any other information that the description of each label. The one-shot example texts were randomly selected for each sample. Both prompts end by providing an output indicator.

3.4 Ranking - spearmans ranking coefficient

When ranking the quality of augmentations in section 4.3, the measure of agreement is the Spearman's Ranking Coefficient. The coefficient represents the agreement between the ranking given by Llama and GPT.

The measure is defined as

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i^2 is the squared difference between the ranks of each pair of observations and n is the total observations. In the experiment $n = 10$, but the rank is not necessarily unique, as multiple texts can have a rank of 0, 5 etc.

$\rho = 1$ indicate perfect agreement, $\rho = -1$ indicate complete disagreement and $\rho = 0$ indicate no correlation between the ranks.

4 Results

In the following section I will present results from three different aspects of text augmentation quality.

4.1 Semantic similarities

Llama and GPT, while similar, are distinct models, leading to variations in their text augmentations. Investigating these differences helps understand the unique approaches of each model. I employ two metrics for this analysis. The first is cosine similarity, which measures how similarly the embeddings of the original and augmented texts are represented in the feature space. The second metric evaluates the lexical diversity by calculating the fraction of tokens shared between the original and augmented texts. This helps determine whether the models are merely replicating the text or introducing new vocabulary.

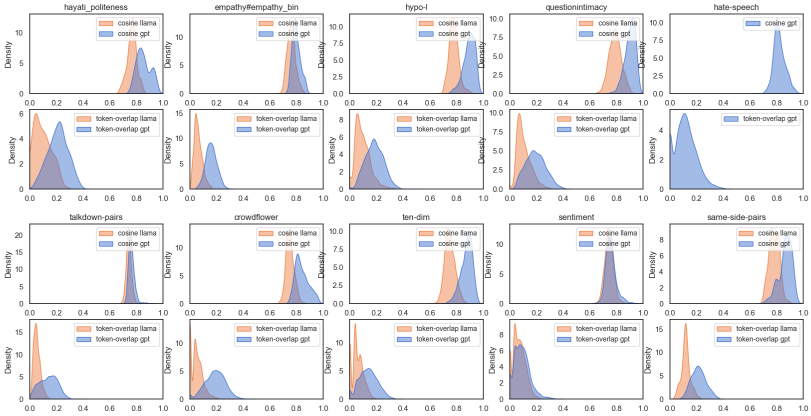

Figure 1: Distribution of similarity metrics between text and augmented by GPT-4 and LLaMA-2 70b. An explanation of the metrics can be in section 3.2

Figure 1 shows the distribution of both metrics for 10 different datasets described in section 2. The figures shows that across all datasets, the augmented data contains substantially different tokens, while still preserving a high cosine similarity. For all the datasets, Llama produce texts, with the least amount of token overlap, naturally leading to a lower cosine similarity. Noticeably, the cosine is relatively high, indicating that even with completely different tokens, the essence of the text is retained. GPT have a higher percentage of token overlaps, but still almost identical to the Llama model. GPT naturally gets a higher cosine similarity score.

4.2 LLM self-assesment

Producing texts, with low token overlap and high cosine similarity is one thing. But does the synthetic texts express the given label to such a degree that each label can be distinguished from each other? In order to address this issue I created the following setup.

The experiment tries to evaluate whether an LLM can detect whether a text belongs to a label given a one-shot text expressing that label. The text I then input to the model is either from that emotion or not. For clarification see the prompt in figure 5 in the appendix. The experiment was done on a sub sample of 100 texts from both the label that the one-shot express and the text that was user input.

Lets first consider the cases where the one-shot emotion and the user inputted label was the same. Figure 2 shows the ratio of times the LLM accurately assess that the text indeed expressed the same emotion.

Generally we see that both GPT and Llama correctly asses that an augmented text in-fact express the label just provided as a one-shot. The scores are generally lower for the GPT-subset indicating that augmentations conducted by GPT are too difficult to asses. Whether this is because the texts are ambiguous in what they express or that the Llama generated augmentations simply are to obvious in what they express is difficult to say.

Interestingly we see that the score for the label *neutral* are very low except when Llama asses the *neutral* augmentations created by itself. As mentioned neutral is the assigned emotion when the human annotators did not agree on what label should be assigned. Therefore the texts are very diverse as a collection, but can individually express a given emotion more

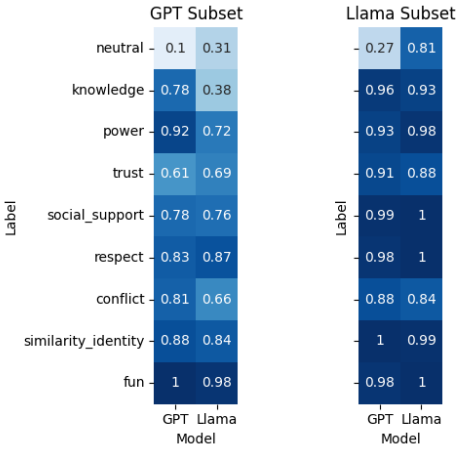

Figure 2: Ratio of times the LLM was able to detect that the user inputted text expressed the same emotion as the text given as an example.

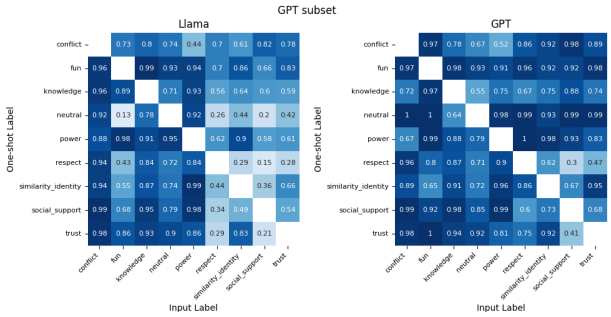

Figure 3: True negative ratio, for augmentations provided by GPT, where the LLM correctly assesed that the label expressed in the user-input was not the same as the label expressed in the one-shot.

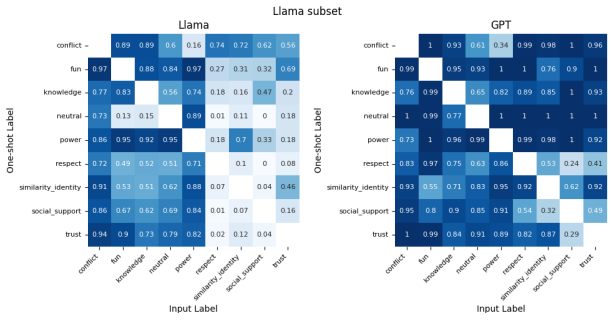

Figure 4: True negative ratio, for augmentations provided by Llama, where the LLM correctly assesed that the label expressed in the user-input was not the same as the label expressed in the one-shot.

clearly than others. In some sense, a low score for this emotion is therefore a good score, since the LLM do not know what to look for. The provided one-shot could for an example express more respect than conflict, but still some of both. As discussed later, labels are not mutually exclusive.

We can know conclude that the LLM performs well in detecting the correct label both in the case of detecting augmentation done by itself or not. But what labels does it struggle to distinguish? Figure 3 and 4 shows the ratio of times the LLM correctly asses that the label expressed in the one-shot was not the same as the label expressed in the user-input.

Immediately we see that GPT on the Llama subset in figure 4, have a few entries where it had a 100% success-rate in detecting that the input label is not the one-shot label. Generally for both subsets, GPT have higher scores, indicating that even for augmented data, not produced by it self, it still are better at distinguishing the labels. For Llama we see that it achieves better performance for augmentations produced by GPT rather than it self. This result could indicate that Llama are better at classifying the label than augmenting them.

In figure 4, we see significantly lower scores for *respect*, *similarity_identity*, *social_support* and *trust*, than the other five. When comparing them to *conflict*, Llama still have a high performance indicating that *conflict* has a distinct characteristics. Same thing is the case when comparing *trust* to *similarity_identity*.

From these results it seems that GPT create augmentations of such a quality that both Llama and GPT are better at classifying those augmentations.

4.3 LLM Emotion Ranking

But does it make sense to compare the classification abilities of Llama and GPT? It could be that the two models, fundamentally have a different perspective on what it means for a text to express a label. For the third and final experiment I ask both LLM's to rank the augmented texts from texts that mostly express a label to expressing a label the least. This is done to asses whether the model fundamentally have the same understanding of each label. This experiment was performed on a small sample of 10 random texts from each of the subsets. The prompt I used to generate the results can be seen in the appendix in figure 6.

Since the spearman coefficient state the level of agreement where 1 state complete agreement and a score of -1 states complete disagreement, we immediately observe that the lowest score the LLM's observe indicate a random ranking for *knowledge* with .04 and .02. When inspecting the ranking of each individual text I observe that the text expressing *knowledge* "*Don't dismiss likes entirely, as they can help you understand if your content is appealing and well-crafted.*" was ranked 9 by GPT but 0 for Llama. GPT finds the text to be expressing a lot of knowledge but Llama do not. When I inspect the reasons GPT made for its decision I see that it argues that "*This text is providing information about the importance of likes in understanding the appeal and quality of content. It is more clearly an exchange of ideas or information, which aligns with the label of knowledge.*" The text with the second highest rank according to Llama, but the second lowest according to GPT is the following: "*Reflecting on my time as a person of color looking for love, it is evident that racism is an ongoing issue in dating.*". The argumentation that Llama provides are the following:"*The text clearly expresses the label "knowledge" because it uses phrases like "reflecting on my time," "looking for love," and "dating." It also mentions racism, indicating that the author has gained experience and understanding from their past encounters, which they are sharing to spread awareness.*". GPT on the other hand states: "*[The text] does provide some insight into the author's personal experiences and observations about racism in dating, which can be seen as a form of knowledge sharing.*". GPT clearly weighs personal experience lower as it says it can be seen as a form of knowledge, where Llama states that the text *clearly expresses the label "knowledge"*, because the author reflects on its own experience.

The models generally agree on the labels *fun*, *social_support* and *conflict*. A *fun* text that both LLM's rank highly are "*LOL, happy to see you having a blast!*", where both model point to "*LOL*" and "*having a blast*". A *conflict* text that both rank highly are "*Why are you so upset if it's not true? Seems like you may be realizing that you do have discriminatory beliefs!*". An interesting case of *conflict* is the following text: "*If you think about it, our world is no stranger to the horrors of sexual mistreatment.*". Llama gives the text a rank of 4, but GPT a rank of 0. Llama stress that the text mention "*horrors of sexual mistreatment*" and that this "*implies a negative and harmful situation*". Again it seems that GPT needs the text more clearly to express the label, where Llama infer its rank based on implications of the text on a more personal note like in the *knowledge* text provided in the previous paragraph.

Emotion	Llama Subset	GPT Subset	Diff
fun	.89	.75	.14
social_support	.79	.82	.03
conflict	.73	.78	.05
similarity_identity	.64	.44	.20
power	.61	.61	.00
trust	.58	.68	.10
respect	.47	.78	.31
neutral	.32	.26	.06
knowledge	.04	.02	.02

Table 2: The Spearman's ranking coefficient score, between the ranking done by Llama and GPT. Values are sorted by Llama subset. The Diff column is the absolute distance between the other two values.

4.4 Discussion and Limitations

When engaging with LLMs characterized by their stochastic nature, it becomes clear that numerous

limitations and discussion points arise.

Firstly, an inadvertent discrepancy in the initial temperature settings was noted between GPT and LLaMA. GPT was initialized with a temperature parameter of 0, whereas Llama was set to 0.7. This discrepancy poses a significant challenge, as the temperature parameter largely determines the model's level of determinism. A cleaner comparison would have been possible if both models operated under identical temperature parameters, thus enhancing the meaningfulness of the comparative analysis.

The methodology employed in the experiments, as detailed in sections 4.2 and 4.3, which involved using subsets of texts from various labels, introduces potential biases. These biases stem from the random sampling of a limited dataset, which could lead to an imbalanced representation of text quality across labels. Notably, the quality disparity is evident in section 4.3, where each label was represented by only 10 texts. These methodological constraints were due to time and financial limitations related to API usage. Expanding the sample size in future experiments would likely enhance the robustness and generalizability of the findings.

Furthermore, the categorization process of the texts under each label assumed a certain standard of quality, with a notable consideration for the non-mutual exclusivity of emotions. During the annotation phase, texts that had disagreements among human annotators were categorized as *neutral* [10]. This approach not only diversified the quality within the *neutral* label but also, indirectly, improved the quality of texts in other categories. However, this raises a significant limitation, particularly apparent in the experiments conducted in section 4.2: it is conceivable for a text to simultaneously convey multiple emotions such as *trust* and *respect*, or *similarity_identity* and *conflict*. The authors of the augmented dataset [10, 11] mention that when annotators disagreed, the text was either replicated into each of the assigned labels or labeled *neutral*. However, the criteria for selecting between these two options were not explicitly detailed in the original authors' methodology.

Lastly, examples of what I term 'context leaks' were found in the augmented texts for Llama. For instance, consider the augmentation from the label *conflict*: "*I strongly disagree with your stance on this issue. It's important to consider all sides before making a decision. (Conveying disagreement and emphasizing the importance of consideration)*". Although the augmentation does not explicitly include the word *conflict*, it concludes by stating that it *conveys disagreement*, which is synonymous with *diverging views*, the definition of *conflict* from table 1.

5 Conclusion

This study has demonstrated the intricate capabilities of large language models, particularly GPT-4 and Llama2 70b, in the realm of data augmentation for texts with subtle social nuances. My investigation revealed that both models are proficient in creating text augmentations with high semantic similarity and lexical diversity. GPT-4 exhibited a stronger capacity for classifying emotions in augmented data, suggesting a possible edge in understanding and replicating nuanced emotional expressions. Conversely, Llama2 70b showcased its strength in producing more diverse augmentations, highlighting its potential in generating varied textual content.

The agreement between the two models in ranking the emotional expression of augmented texts was noticeable, though not without discrepancies. These differences underscore the unique interpretative lenses through which each model views text, thereby influencing their output. This aspect is particularly important when considering the application of these models in sensitive areas where understanding subtle social cues is crucial.

In conclusion, the findings of this study contribute valuable insights into the potential and limitations of LLMs in data augmentation tasks, particularly in handling texts with subtle social distinctions.

Link to the code: `https://github.com/AndersGiovanni/worker_vs_gpt/tree/augmented-evaluation/src/worker_vs_gpt/evaluation`

References

1. Belinkov, Y. & Bisk, Y. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173* (2017).
2. Coulombe, C. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv preprint arXiv:1812.04718* (2018).
3. Niu, T. & Bansal, M. Adversarial oversensitivity and over-stability strategies for dialogue models. *arXiv preprint arXiv:1809.02079* (2018).
4. Kobayashi, S. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201* (2018).
5. Kumar, V., Choudhary, A. & Cho, E. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245* (2020).
6. Sennrich, R., Haddow, B. & Birch, A. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709* (2015).
7. OpenAI. *GPT-4 Technical Report* 2023. arXiv: 2303.08774 [cs.CL].
8. Touvron, H. *et al. LLaMA: Open and Efficient Foundation Language Models* 2023. arXiv: 2302.13971 [cs.CL].
9. Dai, H. *et al. AugGPT: Leveraging ChatGPT for Text Data Augmentation* 2023. arXiv: 2302.13007 [cs.CL].
10. Møller, A. G., Dalsgaard, J. A., Pera, A. & Aiello, L. M. *Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks* 2023. arXiv: 2304.13861 [cs.CL].
11. Møller, A. G., Pera, A. & Aiello, L. M. *The Parrot Dilemma: Human-Labeled vs. LLM-augmented Data in Classification Tasks* unpublished. 2023.
12. Choi, M., Pei, J., Kumar, S., Shu, C. & Jurgens, D. Do LLMs Understand Social Knowledge? Evaluating the Sociability of Large Language Models with SocKET Benchmark. *arXiv preprint arXiv:2305.14938* (2023).
13. Choi, M., Aiello, L. M., Varga, K. Z. & Quercia, D. *Ten social dimensions of conversations and relationships* in *Proceedings of The Web Conference 2020* (2020), 1514–1525.
14. Hayati, S. A., Kang, D. & Ungar, L. Does bert learn as humans perceive? understanding linguistic styles through lexica. *arXiv preprint arXiv:2109.02738* (2021).
15. Buechel, S., Buffone, A., Slaff, B., Ungar, L. & Sedoc, J. Modeling empathy and distress in reaction to news stories. *arXiv preprint arXiv:1808.10399* (2018).
16. Zhang, Y. & Wan, X. MOVER: Mask, overgenerate and rank for hyperbole generation. *arXiv preprint arXiv:2109.07726* (2021).
17. Pei, J. & Jurgens, D. Quantifying intimacy in language. *arXiv preprint arXiv:2011.03020* (2020).
18. Sigurbergsson, G. I. & Derczynski, L. Offensive language and hate speech detection for Danish. *arXiv preprint arXiv:1908.04531* (2019).
19. Wang, Z. & Potts, C. Talkdown: A corpus for condescension detection in context. *arXiv preprint arXiv:1909.11272* (2019).
20. *Sentiment Analysis in Text* https://data.world/crowdflower/sentiment-analysis-in-text. Accessed: 14 December 2023.
21. Rosenthal, S., Farra, N. & Nakov, P. SemEval-2017 task 4: Sentiment analysis in Twitter. *arXiv preprint arXiv:1912.00741* (2019).
22. Körner, E., Wiedemann, G., Hakimi, A. D., Heyer, G. & Potthast, M. *On classifying whether two texts are on the same side of an argument* in *Proceedings of the 2021 conference on empirical methods in natural language processing* (2021), 10130–10138.

23. Vaswani, A. *et al. Attention is all you need* in *Advances in neural information processing systems* (2017), 5998–6008.

24. Wang, L. *et al.* Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv preprint arXiv:2212.03533* (2022).

25. DAIR.AI. *Prompt Engineering Guide* `https://www.promptingguide.ai/`. Accessed: 2023-10-10. 2023.

Appendix

System:

You are an advanced classifying AI. You are going to receive a text written by a user.

Each text expresses one of the following labels: knowledge, power, respect, trust, social_support, similarity_identity, fun, conflict, neutral.

The following is the definitions of the labels:

- knowledge: Exchange of ideas or information,

- power: Having power over the behavior and outcomes of another,

- respect: Conferring status, appreciation, gratitude, or admiration upon another,

- trust: Will of relying on the actions or judgments of another,

- social_support: Giving emotional or practical aid and companionship,

- similarity_identity: Shared interests, motivations, outlooks or Shared sense of belonging to the same community or group,

- fun: Experiencing leisure, laughter, and joy,

- conflict: Contrast or diverging views,

- neutral: neutral communication

User input:

You are tasked with classifying this question: Does the text written by the user express {original_label}?.

An example of a text that expresses {original_label} is: "{original_text}", but the text can vary in many ways and contain completely different words.

You should start your response with a clear yes/no answer. Then in the sentence after, give a short description why you respond the way you do.

User input sentence: {augmented_text}

Answer:

Figure 5: The prompt that was used in section 4.2, when the LLM had to distinguish between the labels.

System:

You are an advanced classifying AI. You are going to receive two texts written by a user.

Both texts expresses one of the following labels: knowledge, power, respect, trust, social_support, similarity_identity, fun, conflict, neutral.

The following is the definitions of the labels:

- knowledge: Exchange of ideas or information,

- power: Having power over the behavior and outcomes of another,

- respect: Conferring status, appreciation, gratitude, or admiration upon another,

- trust: Will of relying on the actions or judgments of another,

- social_support: Giving emotional or practical aid and companionship,

- similarity_identity: Shared interests, motivations, outlooks or Shared sense of belonging to the same community or group,

- fun: Experiencing leisure, laughter, and joy,

- conflict: Contrast or diverging views,

- neutral: neutral communication

Both the texts contain the same label, but the label can be expressed in many ways and contain completely different words.

Your task is to classify which of the two texts most clearly expresses the label. You have to pick one of the two texts. Draw is not an option.

When providing your answer, please answer clearly with "Text 1" or "Text 2" as the first response and then provide a short description of why you respond the way you do.

Examples of good answers are: "Text 1 - [reason]", "Text 2 - [reason]".

User input:

The texts express the label: {label}.

Text 1: {text1.augmented_text}.

Text 2: {text2.augmented_text}.

Which text do you think most clearly expresses the label? Please answer clearly with "Text 1" or "Text 2" as the first response and then provide a short description of why you respond the way you do.

Answer:

Figure 6: The prompt that was used in section 4.3, when the LLM had to rank the quality of augmented texts.