Analyzing Sentiments and Aspects of Fertility Drug Discussions in Social Media: A Comparative Study of Few-Shot Learning and Fine-tuning Techniques with BERT and Large Language Models

Sara El-Beit Shawish

University of Applied Science Darmstadt

Motivation

The current sentiment analysis model used by Merck operates at a document/review level, categorizing entire reviews as either positive or negative. This approach overlooks the nuances of users discussing multiple aspects of a product and expressing different emotions toward them. To address this limitation, the aim is to implement aspect-based sentiment analysis approach, allowing for a more granular analysis of user reviews. Traditional machine learning algorithms struggle with the unstructured and high-dimensional nature of social media text data. Recent advancements in Natural Language Processing (NLP), however, enable the analysis of such data and the extraction of valuable insights. Language models like BERT can be employed for sentiment analysis and to identify specific aspects of fertility drugs, produced by Merck, that express positive or negative emotions. The main challenge lies in creating a domain-specific dataset for supervised training, as existing language models may lack familiarity with certain domain-related nuances and terminology. Thus, there is a need to explore methods for automatically generating high-quality labeled data tailored to this specific sentiment analysis task.

Aspect-based sentiment analysis

Aspect-based sentiment analysis is a text analysis technique focused on discerning sentiments related to specific aspects or features within a given context, enabling a more nuanced understanding of opinions. By dissecting and categorizing sentiments at a granular level, it provides valuable insights into the diverse aspects influencing overall sentiment in user-generated content.

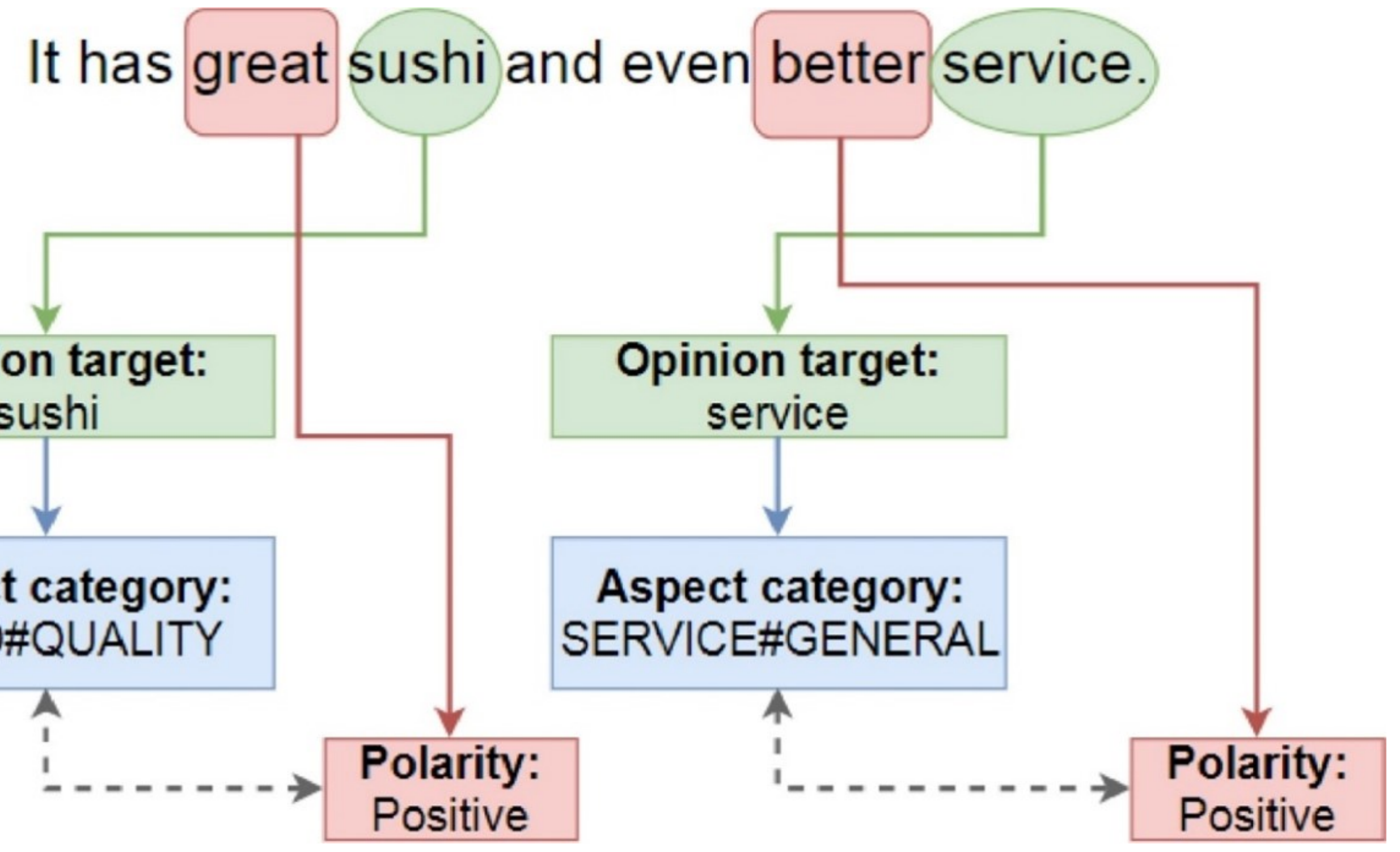
It has great sushi and even better service.
Opinion target: sushi
Aspect category: FOOD#QUALITY
Polarity: Positive
Opinion target: service
Aspect category: SERVICE#GENERAL
Polarity: Positive

Figure 1. Aspect-based sentiment analysis [1]

Objective and Research Questions

The objective of this research is to comprehend patient opinions on specific fertility drugs by extracting and summarizing relevant information from user reviews. Stakeholders prioritize five pre-defined key aspects for analysis: effectiveness, adverse events, availability, administration and costs and the sentiments positive and negative. In order to fine-tune a language model, that can classify those classes, we need to create an annotated dataset. Therefore, two research questions can be derived from this use case:

- Can the use of large language models such as GPT-3.5 turbo help to create a domain-specific dataset and therefore reduce labeling costs?
- Can a mix of human-generated- and GPT-3.5 turbo generated labels further boost performance at a lower cost?

Transformer Model

Transformer models are current state-of-the-art language models that can handle multiple language tasks such as question-answering, text summarization or text classification. They can be classified into three different types:

- Encoder-only These models are typically used to create a representation of the input sentence, primarily for tasks like classification or sequence labeling. Well-known encoder-only models include BERT and RoBERTa.
- Decoder-only Such models are usually used for auto-regressive sequence generation tasks. GPT-3.5 turbo is an example for such a decoder-only model which has shown exemplary results on NLP tasks.
- Both Encoder and Decoder These models are widely employed in sequence-to-sequence generation tasks, such as neural machine translation, where the generation of tokens relies on both the original input and the tokens already generated. T5 and BART are popular examples of encoder-decoder transformers.

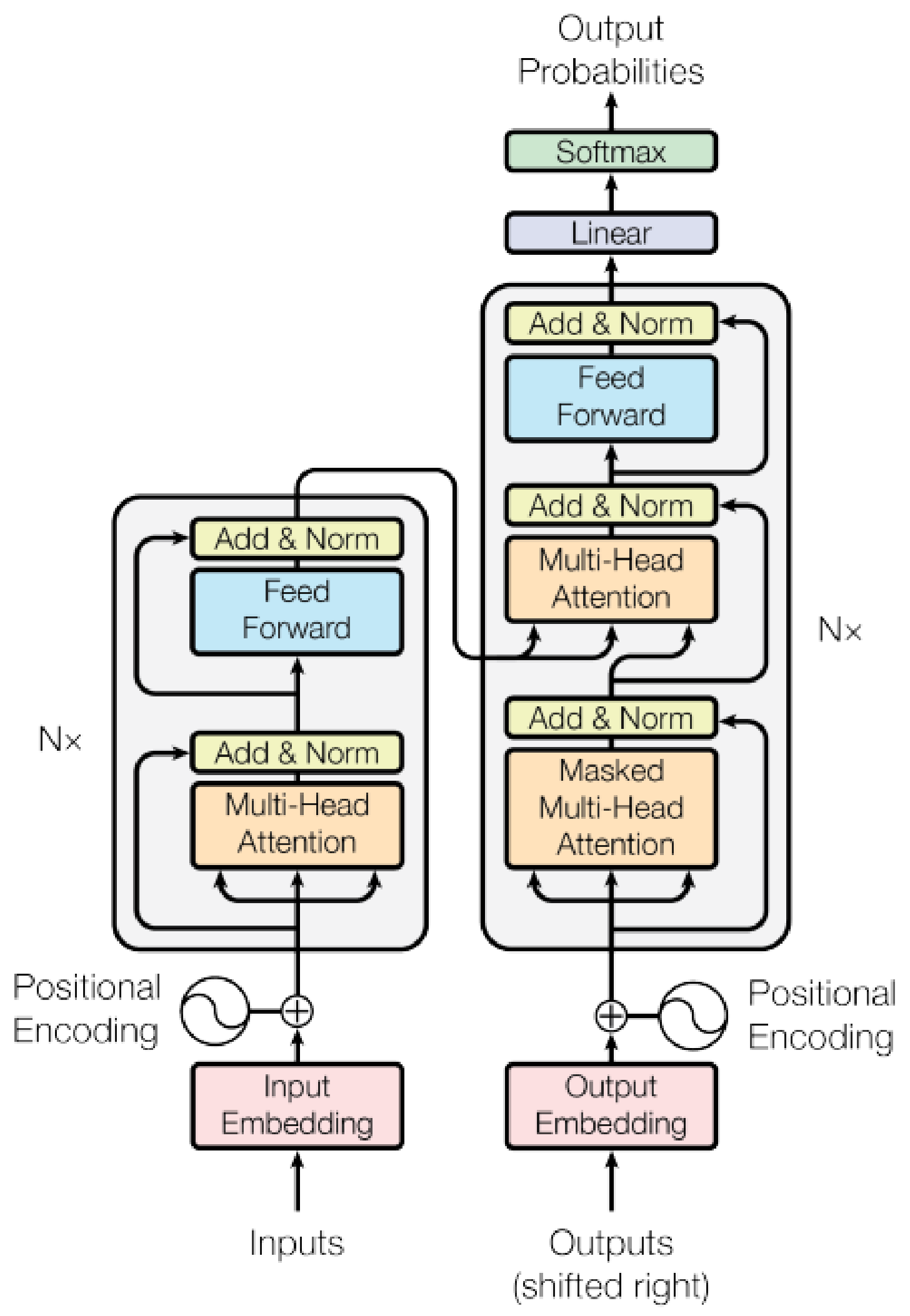

Figure 2. Transformer architecture [2]

Proposed Method

The method contains the following steps:

- Step 1: Creating a synthetic dataset that contains the relevant opinions in the user review and the associated aspect- and sentiment labels.
- Step 2: Fine-tuning two smaller BERT-Model for the aspect and sentiment classification tasks.
- Step 3: For comparison we create a human-annotated dataset and also fine-tune the BERT-Models.

Results

Comparision of all models used in the experiments

Data	Human	GPT-3	Mix Human and GPT-3.5
Test Acc Sentiment	0.92	0.877	0.89
Test Acc Aspect	0.87	0.92	0.95

Table 1. Results for all models used in the experiments.

We are observing, that the aspect prediction task with a BERT model, which was fine-tuned with GPT-3.5 turbo annotated data, achieved a test accuracy of 0.92, which even outperformed the model trained on human data. There are two possible explanations. One of them is, that humans also tend to make mistakes during labeling. Even though we tried to reduce mistakes by defining an annotation guideline, they are not completely avoidable. This 'dirty data', affects the model performance negatively. Another explanation is, that Human performance on a task is not an upper bound on Large Language Mode (LLM) performance considering an LLM has seen much more data compared to a human. In contrast, the sentiment prediction performance was slightly lower with a test accuracy of 0.87, which indicates that the prompt was defined better for the aspects while being vaguer for the sentiment prompts. We get the best result with a mixed dataset consisting of 50% human-annotated data and 50% GPT-3.5 turbo-annotated data resulting in a test acc of 0.95. This implies a potential 50% reduction in human annotation efforts.

Future Work

- Automated Prompt Engineering is very sensitive and changing one word can lead to major changes in the output. The way a user usually deals with this problem is by using a trial-and-error approach and manually adjusting the prompt until it outputs the desired results. It is sometimes unclear what exactly led to those changes and to manually adjust the prompt accordingly. Therefore, it is crucial to investigate methods, that help to create effective prompts in an automated way.
- It can be considered to incorporating an additional sentiment class 'neutral' for an even more detailed analysis
- Extend the size of the training dataset to improve the fine-tuning performance

References

[1] Do, H. H., Prasad, P., Maag, A., and Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: A comparative review. *Expert Systems with Applications*, 118:272–299.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Łukasz Kaiser, and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-December.