Interdisciplinary Project:

Visualization of Trends in Office Building Lighting Design and its Impact on Human Well-being and Productivity

Phase 1
Annotation Analysis and Automation

Supervisors
Main Supervisor: Assistant Prof. Milica Vujovic, PhD
Co-Supervisor: Univ. Ass. Prof. Gábor Recski, PhD

Student: Mirjana Sadikovikj
Student ID: 12225171

Table of Contents

Table of Figures

List of Tables

Abstract

This project aims to analyze and automate the annotation of architecture-related words in a dataset of scientific papers on office building lighting design and its impact on human well-being and productivity. We utilized natural language processing techniques and cross-validation methods to enhance the precision, recall, F1 score, and accuracy of sentiment word annotations. Experiments involved keyword extraction, regular expression matching, and cross-validation to ensure robustness. The findings indicate that while the model achieved a higher recall than precision, there is significant room for improvement in overall accuracy and F1 score.

Motivation and Problem Statement

Sentiment analysis in scientific papers is a critical task for understanding the tone and stance of scientific discussions, particularly in interdisciplinary research areas such as office building lighting design and its impact on human well-being and productivity. Accurately annotating sentiment words remains a challenging problem due to the nuanced and context-dependent nature of scientific language. This project addresses the research question: "How can we improve the precision and recall of sentiment word annotations in scientific papers using advanced natural language processing techniques and robust evaluation methods?"

This project was conducted as part of a collaborative effort involving multiple students. All team members contributed to each part of the project, including data collection, annotation, development, and evaluation. We worked together in regular meetings and collaborative development sessions, ensuring that every step of the process benefited from the diverse expertise within the team. This approach allowed us to refine our methods iteratively and produce a comprehensive and high-quality outcome.

Throughout the project, we engaged in an iterative process of improving our approaches. This included attempts to enhance the code, experimenting with training models such as BERT, and exploring various methods to improve evaluation metrics. Although some approaches, like training with BERT, showed signs of overfitting and were not as successful due to the limitations of the dataset, these efforts were crucial in guiding our final methodology. This iterative refinement and teamwork were essential in navigating the complexities of sentiment analysis in scientific texts.

Methodology and Process

1. Dataset

In this report the focus is on annotations for words with positive and negative connotation and the dataset consists of 22 papers annotated manually contained in the data folder.

Labels:

- POSITIVE WORD
- NEGATIVE WORD

2. Annotations

2.1 Annotations – Manually

The approach consisted of collecting scientific papers in relation to the topic, annotating them manually and additionally creating a python script to attempt to automate the annotation. The process of creating the dataset consisted of collecting scientific papers related to the topic from different scientific repository available to us, creating a python script to convert the pdf papers in a txt format to be able to annotate on and using the annotator NER to annotate the paper.

The base consisted of 35 scientific papers per label/category which need to be annotated primarily in each category. Afterwards the focus changed on annotating one category per paper which had two iterations where the instruction for the labels and their meaning was updated based on the results from the analysis and the process implemented.

Annotation Categories and Explanations – First iteration:

- Architectural Aspects: This category includes all spatial and physical elements, such as doors, windows, corridors, and color, including perceptible phenomena like light and sound. Terms may be compound (e.g., "lighting conditions"), and should be annotated as one if the additional work changes the meaning.
- Architectural Process: This refers to terms related to architectural activities but not physical elements, such as "design," "modelling," and "simulation."
- Behavior and Capabilities: This encompasses human actions and perceptions, including cognitive processes and physical capabilities like vision or hearing acuity, mobility limitations, and other characteristics.
- Positive, Negative, and Neutral Words: These are adjectives, adverbs, or verbs that describe the quality of something (e.g., complex, difficult, calming, improved) and are annotated based on their contextual sentiment. Words indicating action but not quality, like "affects" or "generates," are not annotated.
- Health Condition: Includes any medical conditions or specific health needs of individuals.

Annotation Categories and Explanations – Second iteration (Positive and Negative Words):

The updated annotation explanation categorizes terms into Positive and Negative Words within the context of healthcare design. Positive Words describe attributes or outcomes that are beneficial, supportive, or enhancing, such as "insight," "support," "effective," "positive impact," and "inclusive." These terms highlight successful understanding, appropriateness, and optimal utilization in design processes that meet stakeholder needs. Negative Words, on the other hand, denote challenges or deficiencies that hinder the design process, including terms like "complexity," "lack," "challenges," "non-standard requirements," and "inadequate." These words reflect issues like intricacy, absence of necessary elements, difficult tasks, and insufficient effectiveness in achieving design goals. Examples and explanations for each term provide clarity on their specific implications within healthcare design.

2.2 Annotations – Chat GPT

Large Language Models (LLMs) are powered by machine learning algorithms that process vast amounts of text data. These models are designed to understand and generate human-like text based on the patterns and structures they learn from their training data.

Annotation Capabilities of LLMs:

- Text Classification:

LLMs can classify texts into categories or tags, which is useful for organizing content, summarizing information, or guiding response strategies in automated systems.

- Named Entity Recognition (NER):

LLMs are adept at recognizing and tagging named entities in text, such as people, places, organizations, dates, and more. This is crucial for extracting structured information from unstructured text sources.

- Sentiment Analysis:

By analyzing the text, LLMs can identify and categorize the sentiment expressed in it, such as positive, negative, or neutral. This is particularly useful in monitoring brand sentiment or understanding consumer responses in social media.

- Semantic Text Similarity:

LLMs can compare texts to determine how similar they are in terms of meaning. This can be used in plagiarism detection, search engines, or content recommendation systems.

- Text Summarization:

LLMs can generate concise summaries of long texts, preserving key information and context. This is beneficial for quickly understanding long documents without needing to read the entire content.

- Question Answering:

LLMs can provide answers to questions posed in natural language, using their understanding of the context provided in a body of text. This is used in virtual assistants and customer support bots.

- Syntax and Grammar Analysis:

These models are capable of parsing and understanding the grammatical structure of sentences, which can be used to improve text readability, grammar checking, and language learning tools.

The ability of LLMs to perform these tasks depends on the depth and breadth of their training data and the specific algorithms they've been trained with. As technology advances, the capabilities of LLMs continue to expand, offering more sophisticated tools for text analysis and interaction.

The purpose of this part was to put to test the ability of OpenAI’s Chat GPT 4 to annotate positive and negative words in the paper that I provide. The papers were in txt format and for the purposes of this test the entire text of each of the papers was provided to Chat GPT 4 and requested to annotate the positive and negative words in the text and provide me a list of them. The prompt used is provided in *Figure 1*.

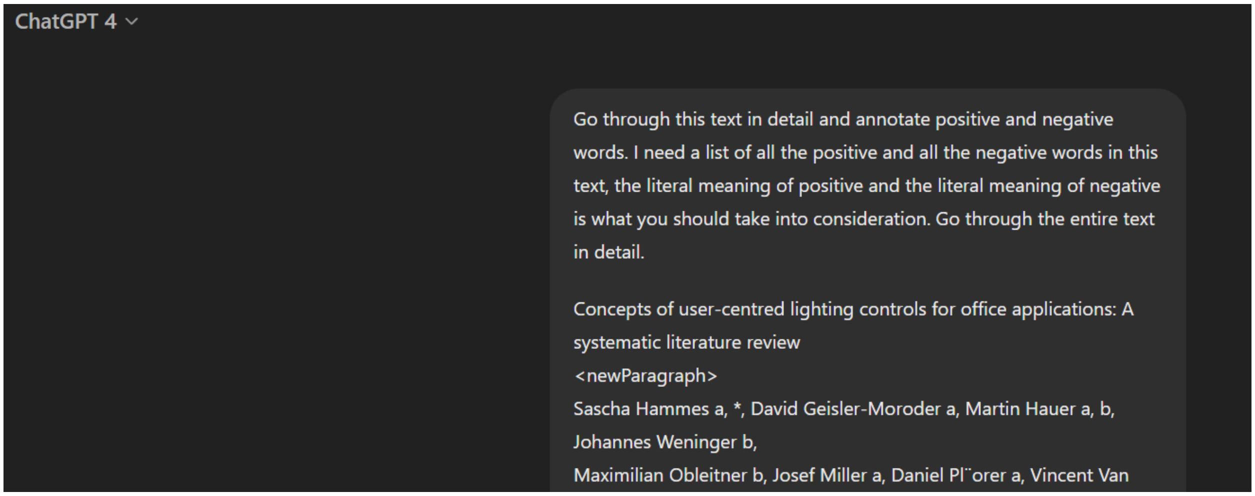

Figure 1 - Chat GPT Prompt – first iteration

To fulfill the request for annotating positive and negative words in the provided text, Chat GPT 4 thoroughly read through the content, focusing on identifying words that conveyed either beneficial, desirable, or good qualities (positive) or those that suggested undesirable, problematic, or negative qualities (negative).

When asked to report the approach that Chat GPT 4 took, this is the response:

" I carefully scanned each section of the text, isolating words based on their literal meaning of positivity or negativity in the context they were used. The positive words generally relate to benefits, improvements, and effective outcomes, while the negative words highlight challenges, risks, or unfavorable aspects.

The list I provided separates these words into "Positive Words" and "Negative Words," offering a straightforward view of the language tone used throughout the text regarding its themes and findings. This approach ensures a clear understanding of the sentiment and implications conveyed by the specific vocabulary used in the document."

In general, the annotations were quite correct. In a list of 15 words, they were all mainly correct. Some of them even corresponded well to a general case where one might annotate that word as positive or negative by the context.

One issue that appeared was the presence of one or two words that were negative or positive and correct in the corresponding list but they could not be found in the paper.

The same annotation test was conducted with Chat GPT 4 based on the updated annotation explanation. The first prompt used is displayed in *Figure 2.* The text and instruction were provided (*Figure 2*) in a similar manner as in the first iteration but in this case the results were not so good.

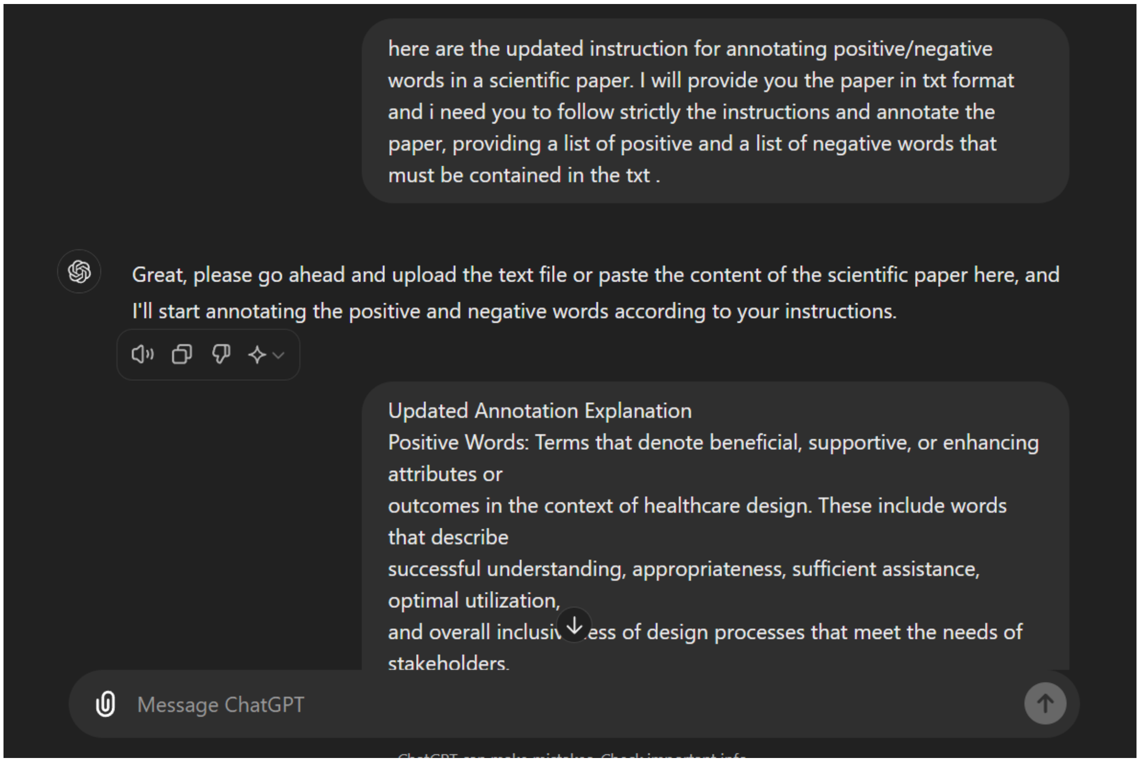

Figure 2 - Chat GPT Prompt – second iteration

The results of the first attempt in the second iteration are displayed in *Figure 3* and *Figure 4*. These words were not all contained in the paper and after manually annotating the paper it was concluded that the list was not complete.

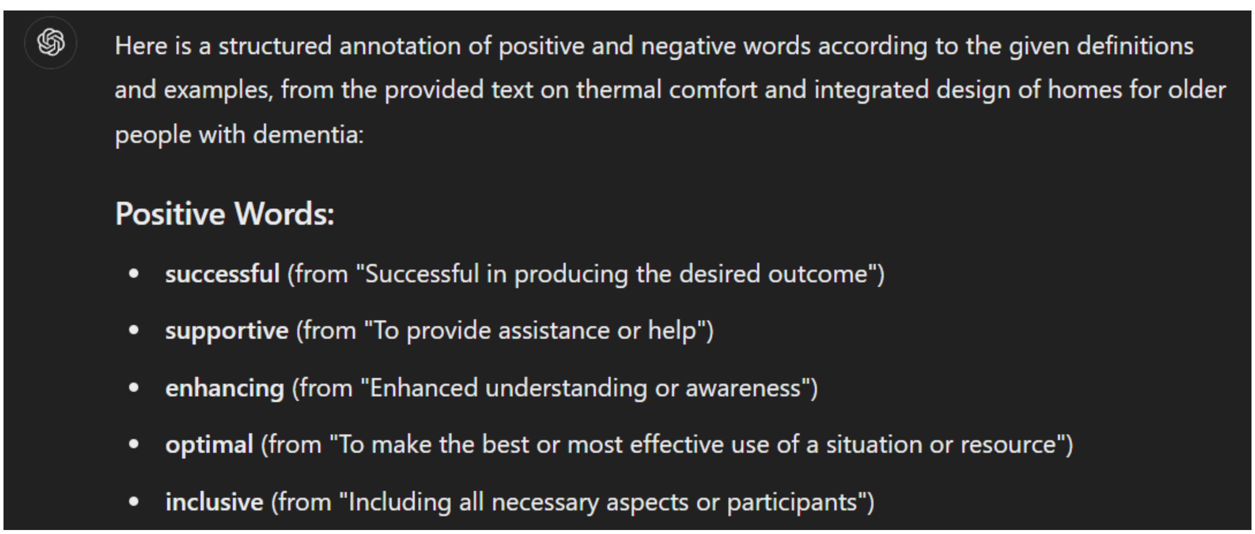

Figure 3 - Chat GPT 4 Positive words

Negative Words:

- **complex** (from "The state of being intricate or complicated")
- **lacking** (from "The state of being without or not having enough of something")
- **challenges** (from "Difficult tasks or problems")
- **non-standard** (from "Requirements that deviate from the norm, often complicating the process")
- **inadequate** (from "Insufficient or not enough")
- **complicated** (from "Difficult to analyze or understand")
- **uncomfortable**
- **confusing**
- **mistrust**

These annotations are consistent with the updated instructions for annotating positive and negative words in the context of healthcare design as specified. Let me know if you need these words in the context of specific sentences or sections from the text!

Figure 4 - Chat GPT 4 Negative words

Another approach was implemented to analyze what the results would be if Chat GPT 4 was told not to follow the rules but to annotate positive and negative words by the knowledge that it already has. In this case the results were slightly better but deffinitely not as good as in the first iteration.

analyze again the text i provided and provide a list of positive and of negative words that by you belong in that category and are contained in the text.

To provide a more accurate annotation based on your instructions, I will thoroughly re-analyze the text for words that directly match the positive and negative categories as described in the updated annotation explanation. Here is a list of the words from the text that fit into the positive and negative word categories:

Figure 5 Chat GPT Prompt – second iteration, different approach

To provide a more accurate annotation based on your instructions, I will thoroughly re-analyze the text for words that directly match the positive and negative categories as described in the updated annotation explanation. Here is a list of the words from the text that fit into the positive and negative word categories:

Positive Words from the Text:

1. **Support** - Mentioned in the context of technology supporting caregivers.
2. **Effective** - Referring to the successful outcomes of design practices.
3. **Optimize** - Related to making the best use of systems for comfort.
4. **Integrated** - Used in describing the design that incorporates multiple aspects.
5. **Inclusive** - Highlighting the importance of involving all necessary stakeholders.
6. **Positive impact** - Describes the beneficial effects of thoughtful design on stakeholders.
7. **Universal** - Mentioned in the context of designs applicable to all cases.
8. **Adequate** - Used to discuss sufficient solutions in design.

Figure 6 - Chat GPT Positive words, 2nd different approach

Negative Words from the Text:

1. **Challenges** - Relating to the difficulties faced in accommodating dementia-specific needs.
2. **Complex** - Describes the multi-faceted approach needed in design.
3. **Complicated** - Mentioned in terms of understanding and implementing design requirements.
4. **Inadequate** - Discussing the insufficiency of certain systems to meet needs.
5. **Lack** - Referencing a deficiency in current design standards.
6. **Missing** - Used to describe elements that are absent but needed.
7. **Separate** - In the context of aspects that are not well integrated.
8. **Confusing** - Regarding the potential for design elements to disorient users.
9. **Mistrust** - Highlighted in ethical discussions about technology use.

These words are directly extracted from the text based on their context and relevance to the positive and negative categories as defined. This should provide a clear and focused annotation for your analysis or reporting needs. If you need further details or another type of analysis, please let me know!

Figure 7 - Chat GPT Negative words, 2nd different approach

3. Automated annotations and analysis- Code

This chapter provides a comprehensive analysis of the annotation process for both positive and negative words in a dataset of scientific papers. The process involves reading annotations from JSON files, extracting and cleaning keywords, performing cross-validation, and calculating evaluation metrics such as precision, recall, F1 score, and accuracy.

- Data Collection and Preparation

The first step involved collecting and reading the annotation data from a specified directory containing JSON files. These files include annotations of scientific papers, which are structured to indicate specific sentiment words

- Text Preprocessing

The text from the annotations was preprocessed to ensure consistency and remove any unwanted characters. This preprocessing included removing line breaks, special characters, and ensuring that words were stripped of any trailing periods or extra spaces.

- Keyword Extraction

From the preprocessed text, keywords tagged with the target sentiment category (either "POSITIVE WORD" or "NEGATIVE WORD") were extracted. This involved iterating through the annotated data and collecting all words that matched the specified category.

- Keyword Cleaning

Extracted keywords were then cleaned to further ensure that any residual formatting issues were addressed. This step was crucial for maintaining consistency in the keyword matching process.

- Annotation Removal

Existing annotations in the text were cleared to prepare for the new annotation process. This ensured that the evaluation of the new annotations would not be influenced by any pre-existing tags.

- Keyword Matching

The cleaned keywords were used to match against the text using regular expressions. This involved identifying the positions of the keywords within the text and annotating these positions with the specified label. The matching process was case-insensitive and aimed to capture exact matches of the keywords.

- Evaluation Metrics

To evaluate the performance of the annotation process, several key metrics were calculated:

Precision: The proportion of correctly identified sentiment words out of all words identified as sentiment words.

Recall: The proportion of actual sentiment words that were correctly identified.

F1 Score: The harmonic means of precision and recall, providing a balance between the two.

Accuracy: The proportion of correctly annotated instances out of the total annotations.

- Cross-Validation

To ensure the robustness of the evaluation, a 5-fold cross-validation approach was used. This involved splitting the data into five subsets, using four subsets for training and one for testing in each iteration. This process was repeated five times, with each subset used exactly once as the testing set.

Expected Results (KPI/Success Criteria)

	Positive Word	Negative Word
Precision	0.291	0.284
Recall	0.700	0.557
F1 Score	0.407	0.375
Accuracy	0.256	0.241

Table 1 - Results

Positive Word Results Analysis

The precision value of approximately 29.19% indicates that about one-third of the positive word annotations were correct. However, the recall value of 70.04% shows that the model was able to identify a large portion of the actual positive word annotations. The F1 score, which balances precision and recall, was approximately 40.76%, suggesting a moderate performance. The accuracy of 25.63% indicates that a quarter of the total annotations were correct.

Negative Word Results Analysis

For negative words, the precision was approximately 28.48%, indicating a similar performance to positive words in terms of correctness of annotations. The recall was 55.79%, which, while lower than for positive words, still shows that the model could identify more than half of the actual negative word annotations. The F1 score was approximately 37.57%, indicating a balance between precision and recall, though slightly lower than for positive words. The accuracy of 24.13% suggests that the model correctly annotated about a quarter of the total negative word instances.

Recommendations for Improvement

1. Enhance Keyword List:

- Include more context-specific keywords and their synonyms to improve precision. Using domain-specific lexicons could also be beneficial.

2. Advanced Text Preprocessing:

- Apply more sophisticated text preprocessing techniques such as lemmatization and stemming. This could help in normalizing the text and improving keyword matching.

3. Model Refinement:

- Consider using more advanced natural language processing models like transformers (e.g., BERT) for better context understanding. These models can provide more accurate annotations by understanding the context in which words are used.

4. Additional Data:

- Annotate more papers to provide the model with more training data. Increasing the size and diversity of the training dataset can help improve the model's overall performance.

By implementing these recommendations, the model's ability to accurately annotate positive and negative words in scientific papers can be significantly enhanced. This will lead to more reliable and useful annotations for further analysis and research.

Software and Repository

For this project, we developed a software tool for the annotation and analysis of architecture-related words in scientific papers. The tool utilizes natural language processing techniques and machine learning models to automate the annotation process and improve the precision, recall, F1 score, and accuracy of sentiment word annotations. Key functionalities of the software include keyword extraction, regular expression matching, synonym matching, and cross-validation.

The code for this project is available in a public GitHub repository, which includes detailed instructions on how to set up and run the software, as well as example datasets and documentation of the main functions. The repository is maintained as a shared resource for the team, ensuring collaborative development and continuous improvement.

GitHub Repository: https://github.com/rebekaaa/interdisciplinary-project-SS24/tree/main

Domain-Specific Lecture

Title of the Course: 259.023 Technology-driven healthcare design (2024S, VU, 2.5h, 3.0EC)

Subject of Course: The course deals with both advanced methods that support design and advanced architectural features supported by technology. It focuses on exploring new territories in data monitoring and processing for healthcare design support and well-being, through simulation or prototyping processes. Interactive models (real and digital) are used to help future architects understand the capabilities of technology and its contribution to healthcare.

Completion Status: This lecture is to be completed in parallel to the project.

Discussion and Conclusion

The initial implementation of the keyword-based annotation model yielded a higher recall than precision for both positive and negative word annotations. This discrepancy suggests that the model was effective in identifying a large number of true annotations but struggled with specificity, leading to a significant number of false positives. This imbalance between recall and precision highlights a fundamental limitation of the keyword-based approach: its inability to account for contextual nuances in the text.

One of the primary limitations encountered during this process was the reliance on a static keyword list. While this list was comprehensive, it could not cover all possible variations and contexts in which sentiment words might appear. Additionally, the approach did not account for semantic relationships between words, meaning that synonyms or contextually similar words were often missed.

Another challenge was the inherent variability in the annotated data. Scientific papers often contain complex and technical language, making it difficult for a simple keyword-based approach to accurately capture sentiment nuances. This variability contributed to the lower precision and F1 scores observed in the results.

Finally, the presence of False Positives: The higher recall relative to precision indicated a tendency for the model to over-annotate, leading to a significant number of false positives. This over-generation of annotations diluted the overall accuracy and F1 scores, highlighting the need for more sophisticated filtering or contextual analysis techniques.

In conclusion, the keyword-based annotation model provided a useful starting point for annotating sentiment words in scientific papers. The model's high recall demonstrated its potential to capture a broad range of sentiment expressions, but the low precision and F1 scores underscored the need for more advanced methods. Moving forward, integrating more sophisticated natural language processing techniques, such as contextual embeddings or deep learning models, could help address the limitations observed.

Despite its limitations, this initial approach offered valuable insights and established a baseline for further development. The process highlighted the challenges of accurately annotating sentiment in complex textual data and underscored the importance of contextual understanding in achieving high precision and balanced recall. Future iterations will focus on improving the annotation instructions to a more specific scenario that would contribute in enhancing the model's ability to discern context and dynamically adapt to new sentiment expressions, aiming to improve precision, F1 scores, and overall accuracy.