Interdisciplinary Project Proposal:

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

Abstract

This project aims to analyze and automate the annotation of positive and negative sentiment words in a dataset of scientific papers related to office building lighting design and its impact on human well-being and productivity. Utilizing a combination of natural language processing techniques and cross-validation methods, the project seeks to enhance the precision, recall, F1 score, and accuracy of sentiment word annotations. The ultimate goal is to develop a robust model that can accurately identify and annotate sentiment words, contributing to more effective sentiment analysis in the context of office building lighting design research.

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?"

Methodology and Process

The methodology for this project follows the CRISP-DM framework to ensure a systematic and reproducible approach. The project involves several key steps: understanding the domain context to define objectives for sentiment analysis, collecting and annotating scientific papers to create a high-quality dataset, and preprocessing text data using NLP techniques for consistency. A keyword-based annotation model will be developed and enhanced with synonym expansion to improve coverage, with robust evaluation ensured through cross-validation. Evaluation metrics such as precision, recall, F1 score, and accuracy will be calculated to assess model performance. The entire process will be documented for reproducibility, and the model will be prepared for deployment in sentiment analysis tasks. Specific methods from the Data Science curriculum, including NLP, experiment design, and reproducibility practices, will be applied throughout the project.

Annotation Categories and Explanations

- Architectural Aspects: This category includes all spatial and physical elements, such as doors, windows, corridors, and color, including perceptible phenomena like light and sound.

- 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.

- Health Condition: Includes any medical conditions or specific health needs of individuals.

Expected Results (KPI/Success Criteria)

The primary objectives of this project are to achieve significant improvements in precision, recall, F1 score, and accuracy for sentiment word annotations. Specifically, the target is to increase precision for both positive and negative words to above 70-80%, achieve recall rates above 70-80%, and aim for an F1 score of at least 70-80%. Additionally, the project seeks to improve overall accuracy to above 85-90%. These targets represent the proportion of correctly identified sentiment words, the proportion of actual sentiment words correctly identified, the harmonic mean of precision and recall, and the proportion of correctly annotated instances out of the total annotations, respectively. The approach involves keyword extraction and synonym matching to enhance the annotation process, with baseline metrics derived from initial model runs providing reference points for evaluating improvements. Through this structured methodology, the project aims to significantly advance the accuracy and reliability of sentiment word annotations in scientific papers.

Approaches and Baseline

The approach involves enhancing the annotation process through keyword extraction and synonym matching. Baseline metrics, derived from initial model runs, will serve as reference points for evaluating improvements. The process includes generating a list of keywords from annotated data, enhancing the keyword list with synonyms for better coverage, and implementing 5-fold cross-validation for robust model evaluation. Annotation and evaluation of the text are performed to calculate metrics assessing the model's accuracy and effectiveness. By following this structured methodology and targeting specific KPIs, the project aims to create a baseline and then significantly improve the accuracy and reliability of sentiment word annotations in scientific papers.

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.

Conclusion

This project aims to contribute significantly to the field of sentiment analysis in scientific literature by developing a robust, accurate, and reproducible annotation model. The methodologies and insights gained will be valuable for further research and practical applications in data science. By implementing the recommendations provided, 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.