Example projects from HT2025

Marcel Bollmann

This document contains *selected* titles and abstracts from project reports submitted in HT2025, in arbitrary order. Only projects that received a passing grade are included.

Analysis of the development of the sentiment in the context of Russia in the German Bundestag

This project performs a sentiment analysis of parliamentarians' views on Russia in the German parliament since 2010. A pre-trained BERT-based sentiment classifier for sentiment analysis is finetuned, and a random forest is trained on a labeled dataset containing 445 speeches in the German language. The random forest shows better performance and is therefore used to evaluate four predefined hypotheses on how sentiment has developed due to events such as the invasion of Crimea in 2014 or the attack on Ukraine in 2022. The results show no support for an increase in negative sentiment due to these events.

Event-Driven Topic Shifts in The Guardian: A Comparative Analysis of BERTopic and LDA around Queen Elizabeth II's Passing

We analyze how The Guardian's news agenda shifted around Queen Elizabeth II's passing, using topic modelling on articles from 2020–2025. To isolate indirect effects, we exclude direct event coverage via buffered windows and keyword filtering. We compare Latent Dirichlet Allocation (LDA) and BERTopic, employing a novel semantic alignment strategy to track evolving themes across independent models. Results reveal a significant "agenda reset": diffuse reporting declined while specific narratives—notably Economic Policy and Political Leadership—consolidated. BERTopic provided superior granularity in capturing these shifts compared to LDA. Our findings demonstrate that major events drive substantial structural changes in media agendas, with established themes briefly displaced before returning with altered intensity.

Linguistic Fingerprints on the Ice: Differentiating Coach and Player Discourse in NHL Interviews

This study investigates the linguistic and thematic distinctions between National Hockey League (NHL) Coaches and Players during Stanley Cup Final interviews (1997–2019). While sports interviews are often dismissed as cliché-filled, we hypothesized that the distinct pressures of the roles; performance execution (Players) versus strategic management (Coaches); would create detectable linguistic signatures. Using a dataset of 2,096 interview transcripts, we trained supervised classifiers reaching an exceptional 92% accuracy. Moving beyond prediction, we interpret our findings using LIME, Log-Odds Ratio, and validate the results with unsupervised clustering. We find that players utilize a "subjective" dialect characterized by first-person pronouns and sensory vocabulary, while coaches employ a "managerial" dialect focused on collective nouns ("the group") and abstract processes. Furthermore, we demonstrate that unsupervised clustering and topic modeling only successfully separates these roles when the dataset is balanced, mitigating the noise of the dominant player class.

Synthetic-data training for Swedish clinical text de-identification (PII NER)

De-identifying clinical text is required for secondary data use in healthcare, but developing automated systems is limited by the scarcity of open training data. This study presents a methodology for training Swedish clinical Named Entity Recognition (NER) models using purely synthetic data. We generated a corpus of 250 structured clinical notes enriched with Personally Identifying Information (PII) using a large language model. We then fine-tuned KB-BERT, a Swedish Transformer model, to detect 9 entity types including both general identifiers (e.g., names, locations) and specific PII (e.g., occupations, healthcare identifiers). We also demonstrated that a sliding window approach for long documents improved recall by 1.5% and reduced PII leakage by nearly 50% compared to truncation. The final model significantly outperformed rule-based and general-purpose NER baselines (spaCy, stock KB-BERT) on clinical PII, achieving a macro F1-score of 0.963 and a leakage rate of 1.7%, compared to >15% leakage for baselines. These results validate the use of synthetic data for training high-recall clinical de-identification systems.

Beyond Keywords: A Hybrid Neural Semantic Approach for Coffee Recommendation

This project implements a hybrid recommendation engine to address the disconnect between informal user preferences and professional coffee cupping notes. Our approach combines regular expressions to enforce objective constraints—such as price, origin, and roast level—with a Sentence-BERT (SBERT) layer for semantic flavor mapping. We evaluated the system across three realistic scenarios: integrated constraints, lexical noise (typos), and complex sensory descriptions. Experimental results show that SBERT outperforms a TF-IDF baseline, particularly in descriptive queries where it achieved a 187.48% improvement in similarity scores by correctly mapping metaphorical intent (e.g., “like a garden”) to specific floral descriptors. The results show that the sharp increase in retrieval confidence did not result in a corresponding jump in BERTScore. This indicates that capturing the underlying sensory profile is more essential for coffee recommendation than the literal token overlap tracked by BERTScore.

Specialist vs. Generalist: Evaluating Aspect-Based Sentiment Analysis in the Skincare Domain

This project investigates the trade-off between specialist efficiency and generalist reasoning in Aspect-Based Sentiment Analysis (ABSA) within the complex, implicit-heavy skincare domain. We try to find the best option when you want to perform an ABSA task but training a domain specific model is not feasible. We compare a fine-tuned specialist ABSA model (DeBERTa-v3-large) against a generalist zero-shot Large Language Model (Llama 3.1 8B) on a custom dataset of Amazon consumer reviews. Evaluating for Aspect Sentiment Classification (ASC), we find that the Llama 3.1 significantly outperforms the specialist baseline, achieving 93% accuracy compared to DeBERTa’s 75%. The LLM demonstrates superior capability in detecting implicit negative sentiment (e.g., “tightness” as a skin reaction) where the dedicated model relies more heavily on explicit lexical aspects. However, this reasoning capability comes at a substantial computational cost, with Llama 3.1 suffering a 55x reduction in throughput (2.6 inferences/sec) compared to DeBERTa (144 inferences/sec).

Unsupervised Discovery and Hierarchical Clustering of Spanish Economic Activities through Deep Semantic Embeddings and Manifold Learning

This project addresses the challenge of automatically identifying and grouping Spanish economic activities from unstructured corporate descriptions known as the "objeto social." Using a dataset of 16,793 manufacturing companies, we propose an unsupervised discovery pipeline that integrates Information Extraction (IE), deep semantic embeddings via the Spanish-specific BETO model, and manifold learning using UMAP. Clustering is performed via HDBSCAN to identify latent industrial structures without a priori assumptions about the number of categories. Results are evaluated against the official CNAE-2009 taxonomy. Our findings demonstrate that this approach effectively recovers the hierarchical organization of the Spanish manufacturing sector while identifying emerging sub-industrial trends, providing a scalable solution for modernizing administrative statistics and corporate intelligence.

Swedish Sentence-Level CEFR Classification with LLM Annotations

The lack of a Swedish sentence-level CEFR dataset hinders readability assessment and language learning applications. To address this, we attempt using large language models (LLMs) as cost-effective annotators. Through pilot studies on 250 sentences, we demonstrate that top-tier LLMs achieve high internal agreement (Quadratic Weighted Kappa 75-92%) and can reliably assess sentence difficulty. We then create the first Swedish CEFR sentence dataset by annotating 10k sentences from COCTAILL, 8 Sidor, and SUC3 using selected models. On this dataset, we evaluate three classification approaches: prototype-based metric learning, BERT fine-tuning, and feature-based classifiers. The prototype-based method with Swedish BERT achieves the best performance with 84.1% macro-F1 and 94.6% QWK, outperforming BERT baseline by 12.1% in macro-F1.

Uncovering Latent Topic Patterns in Untappd Reviews using BERTopic

This study examines consumer description variations across major beer styles by uncovering latent topic patterns of beer reviews from Untappd using BERTopic. Following preprocessing to remove "social noise," the model was refined through a iterative-optimization processes. This process isolated style-defining topics that accurately capture distinct flavor profiles, such as "roasty" for Stouts and "raisin" or "toffee" for Barleywines. Results demonstrate a clear alignment between specific topics and beer styles, as each style is defined by its own unique and consistent set of words. By

isolating flavor descriptors from general lifestyle contexts, this research bridges the gap between technical descriptions and authentic consumer language, transforming unfiltered social feedback into actual flavor profiles.

What makes a Kickstarter project successful?

The goal of this study has been to understand what words in the description of a Kickstarter project make the project successful. The study started by collecting data on Kickstarter projects from Webrobots.io and scraping project descriptions from the Kickstarter website. This data was then used to fine-tune a BERT model that would predict if a project was successful or not, based on the description. The LIME algorithm was implemented to explain what words impacted the model's predictions. The BERT model showed a 79% accuracy in predicting the success of a project. The results from the LIME algorithm showed that naming credible organizations, such as Berklee College of Music or Sundance Film Festival, can have an impact on the success of the project.

Identifying Large Language Models from Generated Text: A Case Study in Italian Counseling Dialogues

We investigate whether Large Language Models (LLMs) can be identified from text alone in controlled Italian counseling dialogues. Using TF–IDF n-gram representations, unsupervised clustering and linear SVM classification achieve near-perfect separation between four LLMs. Interpretability analysis shows that lexical choices, phrasal patterns, and even personal names contribute to model fingerprints. Simple formatting features alone reach 90% accuracy, indicating that surface-level cues also reflect model-specific tendencies. These results demonstrate that LLM-generated texts carry distinctive, detectable signatures under identical prompts and dialogue structures.

Quantifying Style: The Evolution of Themes and Writing Styles in Science Fiction

Since the creation of science fiction, authors have sought to make meaningful stylistic and conceptual contributions to a genre limited only by imagination and vague notions of technological feasibility. This paper seeks to analyze the evolution of science fiction from the 19th century to the modern day by means of a computational, NLP-based approach. Looking at a collection of novels by six leading authors in the genre (Shelley, Verne, Wells, Asimov, Vonnegut, and Doctorow), the text is processed, the

features extracted, and keywords tracked to compare each author's style of writing and how the topics discussed have shifted with the advent of new technology. The results indicate that 19th-century prose was characterized by its descriptiveness and necessary invention of new terms, while the 20th and 21st-century science fiction authors adopted a more casual tone, showcasing human interactions just as much as the fantastical and bizarre. A biological worldview was forced to adapt to the creation of advanced machines and the consequences that might have on society, although concrete shifts are difficult to discern. A Random Forest classifier achieved an accuracy of 75%, identifying average word and sentence length as the primary determining factors of writing style and genre evolution.

Comparative Analysis of Retrieval-Augmented Generation Architectures for Financial Question Answering

This study evaluates five Retrieval-Augmented Generation (RAG) architectures for financial QA using the FiQA dataset. We compare Naive RAG, Re-Ranking RAG, Hypothetical Document Embeddings (HyDE), Query Expansion RAG, and Multi-Step Iterative RAG across lexical, semantic, and retrieval metrics. Results show Re-Ranking RAG achieves best lexical performance (8% higher ROUGE-L, 177% higher BLEU vs Naive) but with higher variability. Naive RAG provides superior semantic preservation (5% higher cosine similarity) and consistency. Retrieval quality shows limited correlation with generation performance ($r < 0.30$), challenging RAG design assumptions.