Example projects from HT2024

Marcel Bollmann

This document contains *selected* titles and abstracts from project reports submitted in HT2024, in arbitrary order. Only projects that received a passing grade are included. The projects are sorted alphabetically by their title and given a `badge` to indicate roughly which area(s) of Text Mining they correspond to.

A Comparative Study on the Impact of Custom Sentiment Features on Spam Email Classification

`Text Classification`

This study explores the impact of incorporating custom sentiment features into spam email classification using three models: Naive Bayes (NB), Random Forest (RF), and Multi-Layer Perceptron (MLP). Each model was evaluated on datasets with and without sentiment features, with performance compared across Precision, Recall, F1 Score, Accuracy, and error metrics. The inclusion of sentiment features improved NB's precision by 1% and reduced false positives by 13.3%, but increased false negatives by 50%. MLP demonstrated a 1% recall improvement and a 12.8% reduction in false negatives, at the cost of lower precision and higher false positives. RF showed minimal changes, reflecting its robustness. These findings highlight that custom sentiment features can enhance specific aspects of spam classification, depending on the model and application priorities.

Topic Modelling of Bike Touring Blogs

`Topic Modelling`

Bicycle touring is a holiday style that can be enjoyed in many different forms. This paper explores the application of topic modelling methods, specifically Latent Dirichlet Allocation (LDA) and BERTopic, to analyze a collection of 2,967 bike touring blogs. The

study aims to uncover patterns in touring styles such as sleeping styles (e.g. camping versus hotels), bike types (e.g. e-bike, mountain bike), and route choices (e.g. road versus off-road) by identifying topics between the blogs. Preprocessing steps include language filtering, named entity recognition (NER), as well as other standard techniques. Results indicate that the models struggle to capture the hypothesized patterns effectively, possibly because the blogs contain much other information regarding the day-to-day. Further work is needed to explore other methods for further refining the preprocessing steps and modelling techniques.

From Questions to Queries: Comparing Text-to-SQL Retrievers for RAG Systems Using Self-Annotated Questions

Information Retrieval

This project evaluates the fine-tuning of text-to-SQL models for enhancing database query generation in Retrieval-Augmented Generation (RAG) systems, using the WHO Life Expectancy dataset. A pre-tuned Flan-T5 retriever was compared to an extended version further fine-tuned on self-annotated questions. The meta-llama/Llama-3.2-1B is used as the generator for the system. Assessments were conducted using execution accuracy, Exact Matching, ROUGE-2 scores and human evaluations. Marginal improvements in exact matching were observed for the retriever, human evaluations on generated outputs are considered as good despite low automated ROUGE-2 scores. The models face challenges with semantic misinterpretations and handling complex queries. Suggestions include expanding training datasets, diversifying annotators, and refining hyperparameters to improve performance. This work highlights the potential of text-to-SQL models to simplify database access for non-technical users while identifying areas for improvement.

Connecting Sentiment Analysis with Player Performance: Insights from Premier League Match Reports

Text Classification

This project investigates the relationship between sentiment in Premier League match reports and player ratings from FotMob. Using sentiment analysis with a pre-trained CardiffNLP Twitter-RoBERTa model, the study analyzes 158 match reports and correlates sentiment values assigned to players with their objective ratings. Sentences mentioning a single player were analyzed to assign sentiment scores, which were then aggregated to get an overall sentiment label for each player. The results reveal a

clear trend: players with positive sentiment tend to have higher FotMob rating, while those with negative sentiment show significantly lower ratings.

Analyzing Parliamentary Discussions in Sweden (2014–2023): A Text Mining Approach Using BERTopic

Topic Modelling

This study employs BERTopic to analyze parliamentary discussions in Sweden from 2014 to 2023. By categorizing discussion points into broader areas, such as healthcare, immigration, and economics, this research evaluates how effectively BERTopic captures trends in parliamentary focus and aligns them with real-world events. A comparative analysis was conducted to assess the impact of the preprocessing steps that is stop word removal and lemmatization—against raw, unprocessed data. Results demonstrate that preprocessing enhances the clarity and interpretability of key topics, with trends such as spikes in healthcare discussions during the COVID-19 pandemic and immigration-related debates during the Syrian refugee crisis. However, significant patterns, like the 2015 Syrian refugee crisis, were detected even without preprocessing. Despite challenges with short text inputs and the absence of labeled data, the findings underscore the value of BERTopic for studying temporal shifts in political discourse.

Evaluating Text Mining Models Ability to Predict Music Genre Based on Song Lyrics

Text Classification

When you hear a song for the first time, it often does not take long before you can determine: “this is hiphop” or “that’s country”. Often many aspects such as tempo, instruments, and rhythm come into play, indicating a certain genre. However, in this paper the genre is predicted by only using the song’s lyrics as input to various classifiers and a large language model. The classifiers include Multinomial Naïve Bayes, Multinomial Logistic Regression, Decision Trees, Random Forests, Gradient Boosting, K-Neighbors and Voting, along with Meta’s LLM, Llama-3.1-8B-Instruct-GGUF. Using sklearn’s count vectorizer for the classifiers and the full lyrics for the LLM, the genre predictors were evaluated and compared. The optimized Multinomial Naïve Bayes reached a precision of 60%, recall of 59% and the highest F1-score of any model reaching 59%. This score was not far ahead of the LLM getting higher precision of 62%, however only 53% recall and 54% in F1-score. Considering the LLM’s complexity and

computational requirements, the Multinomial Naïve Bayes classifier proves superior, being both simpler and significantly faster.

Your honor, I RAG my case–Retrieval Accuracy Exploration for Separated Databases on Law Text Data

Information Retrieval

The paper concerns the emergence of generative AI solutions, and how to apply them in the legal sphere. Specifically, a Retrieval Augmented Generation (RAG) system is implemented using two different methods. One baseline method where all of the data is stored in a single vector database, and one where the different types case outcomes is stored in a different vector database. The goal of the paper is to understand if legal practitioners can benefit from annotating their data before adding them to the knowledge base that is used for a potential RAG system. The results show that, for this dataset and retrieval set-up, the separate database method has statistically significant positive effect on the accuracy at $k = 1$ metric, but not statistically significant for accuracy at $k = 5$. Even though the set-up is not optimized and there are several areas of improvement discussed, the results indicate that legal data similar to the data used should be annotated and stored in separated databases for increased accuracy at k. However, further experimentation is needed.

Code Generation Capabilities of LLMs: A Comparative Analysis Using Competitive Programming Problems

Text Generation

Large Language models (LLMs) have demonstrated remarkable advancements in natural language processing, but their performance on competitive programming tasks remains an open question as benchmarks usually focus on more traditional coding tasks. Competitive programming tasks require not only linguistic understanding but also algorithmic and mathematical reasoning, precision and efficiency. This paper evaluates the problem-solving capabilities of 5 small size LLMs on a set of competitive programming problems from the publicly available coding platform Open.Kattis. Each code generated is evaluated on compiling correctness, success on sample input/output combinations and finally on the online judging platform directly. The results reveal a significant gap between the ability of the LLMs to produce correct code, which is pretty good, and the ability of the LLMs to produce code that actually solves the competitive programming tasks, which is not very good. These finding offer a valuable insight to the current state of most LLMs, which is that although they

appear to produce valuable outputs, they actually lack strong reasoning capabilities to produce qualitative outputs.

Evolving Corporate Strategies and Technological Innovation: A Topic Modeling Analysis of Form 10-Ks from Google, Microsoft, Apple and Meta

Topic Modelling

This study employs topic modeling techniques to analyze 10-K filings from 2017 to 2023 for four major technology companies: Google, Microsoft, Apple and Meta. The primary objective is to examine how the topics discussed in these forms evolved around external events such as the COVID-19 pandemic and to evaluate the growing prominence of Artificial Intelligence (AI) in corporate strategies. Two topic modeling approaches were used—Latent Dirichlet Allocation (LDA) and BERTopic. The dataset comprising text segments from annual filings was curated and pre-processed extensively. The analysis confirmed that the 10-K topics significantly shifted during the pandemic, reflecting companies' strategic adaptations to the global crisis. Additionally, AI emerged as an increasingly discussed theme, with a notable rise in prominence by 2023.

Comparative Analysis of DistilBERT and Traditional Classifiers for Multi-Class News Classification

Text Classification

Text classification is a crucial subfield of Natural Language Processing (NLP), aiming to categorize textual data into predefined categories. This project compares the performance of traditional machine learning methods, specifically the Naive Bayes classifier and Logistic Regression, with the more advanced DistilBERT model. Hyperparameter tuning is performed on the DistilBERT model, focusing on hidden size, dropout rate, and the number of epochs. The AG News dataset is used for this study, with subsamples of sizes 400, 1000, 5000, and 20000 to train the models. Results show that DistilBERT consistently outperforms traditional methods in all configurations. However, the gap in accuracy is notably larger when working with smaller datasets, highlighting that pre-trained model can be more useful with small training data sizes.

Linguistic and Semantic Strategies in Fake News

Topic Modelling | Text Classification

This study utilizes NLP techniques and machine learning models to investigate the linguistic and semantic differences between fake and true news articles. By analyzing a balanced dataset of fake and true news articles, this research employs sentiment analysis, topic modeling, and bias detection to uncover key distinctions. Sentiment analysis reveals that fake news tends to exhibit more extreme emotional tones, leveraging both positive and negative emotions to attract readers, whereas true news maintains a neutral tone aligned with factual reporting standards. Topic modeling identifies five recurring themes, highlighting speculative and emotionally charged language in fake news, while true news emphasizes structured and factual narratives. Bias detection shows that fake news articles demonstrate higher proportions of both positive and negative bias, contrasting with the predominantly neutral tone of true news. These findings underscore the value of integrating sentiment, thematic, and bias analyses into automated fake news detection systems. These methods not only improve detection accuracy but also provide deeper insights into the linguistic strategies employed in misinformation campaigns. Future research should incorporate multi-lingual datasets to address evolving misinformation strategies and mitigate their societal impacts.

Analyzing Gender and Age Differences in Symptom Descriptions from Reddit Posts in the 'AskDocs' Subreddit

Information Extraction

This project investigates trends in symptom description from posts in the online forum 'AskDocs' for different gender and age groups by using natural language processing techniques. The symptoms are extracted from the posts by using Named Entity Recognition (NER) with the biomedical SpaCy model, en_core_sci_sm. The entites are linked to medical concepts in the Unified Medical Language System (UMLS) through ScispaCy. The analysis categorizes symptoms by gender and age, identifying the most frequently mentioned symptoms and the results are visualized using bar plots. The results indicates that there are more females than males using this forum to ask for medical advice and that the most active age group is people between 19 to 30 years old. The most common symptom is pain and we can see that pain is reported more by females than by males which aligns with existing research on gender differences in chronic pain. The project has some limitations, such as a limited and biased dataset and the use of a relatively small NLP model. Addressing these aspects could lead to a more robust analysis.

Evaluating and fine-tuning Instruction-Tuned LLMs for Complex Question Answering Tasks

Text Generation

This study evaluates the performance of instruction-tuned Large Language Models (LLMs) with parameters ranging from 1 to 8 billion, on multiple choice question answering tasks. The dataset used for this study is the RACE dataset, which contains reading comprehension questions from Chinese English exams. The study evaluates six different models including Llama 3, Mistral, Falcon, and DeepSeek LLM variants. The results revealed that the Falcon 7B Instruct model achieves the highest accuracy at 82.0%, while the smallest model, Llama 3.2 1B Instruct achieved the lowest accuracy of 38.5%. By using PEFT along with LoRA, it is shown that a Llama 3.2 3B Instruct model can be optimized to achieve 79.5% accuracy, surpassing the accuracy of the Llama 3.1 8B Instruct model for this task. The findings of this study suggest that there is some correlation between model size and performance, but a larger model might not be necessary for all tasks, and fine-tuning a smaller model can significantly enhance its capabilities.

Can sentiment analysis predict economic crashes?

Text Classification

The purpose of this project is to analyze and compare the sentiments of the economy in the years leading up to the financial crisis of 2008 to the current economy. This is achieved by utilizing a large language model fine-tuned on financial news to perform sentiment analysis on financial news article titles and summaries. The sentiment of the economy was largely negative during the year of 2007, but seemed to recover at the start of 2008 before continuing to deteriorate during the lead-up to the climax of the crisis. The trend produced for the years 2007-2008 in this project matches the trend produced by a numeric approach to estimating market sentiment of those years. The current sentiment of the economy is roughly half as negative as the peak seen during 2007.