Universiteit Leiden

Master Computer Science

Issue Detection and Future Proofing Dutch Government Apps Using Language Technologies

Name:	Anca-Mihaela Matei
Student ID:	s4004507
Date:	2025-05-18
Specialisation:	Artificial Intelligence
1st supervisor:	Natalia Amat Lefort
2nd supervisor:	Flor Miriam Plaza del Arco

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

As public services increasingly shift to digital platforms due to e-Government initiatives, understanding and incorporating user feedback has become critical for improving the quality and usability of government applications. The field of Natural Language Processing (NLP) has emerged as a crucial response to the need for processing and analyzing vast and diverse user feedback, offering techniques for extracting meaningful insights from human language. Among these techniques, Large Language Models (LLMs) have become key scalable and versatile tools. They can perform a wide range of tasks, such as summarization, instruction following, and classification, without the need for extensive input preprocessing. Building on these capabilities, this thesis explores the application of LLMs to extract, classify, and forecast issues reported in user reviews from four Dutch government applications, namely KopieID, Reisapp, MijnOverheid, and DigiD. This research thesis is structured around four core tasks: (1) issue extraction, (2) multi-label review classification, (3) assessment of how different issues impact star ratings, including a temporal analysis, and (4) forecasting of future issues and actionable recommendations. A comparative analysis between LLMs and Latent Dirichlet Allocation (LDA) is performed to evaluate coherence and classification confidence (via Shannon Entropy). The results show that LLMs outperform LDA in coherence, flexibility, and interpretability, though challenges such as hallucination and classification ambiguity were observed. The star-rating assessment highlights that technical reliability remains a key driver of user dissatisfaction, while usability-related concerns exhibit more variable effects across applications. Forecasting analysis reveals that LLMs can partially identify emerging issues and generate precise, app-specific recommendations, though the prediction of issues' frequency remains limited. This research offers a replicable, unsupervised pipeline for multilingual user feedback analysis and provides practical insights for enhancing citizen-centric digital services in the public sector. Government institutions could use and built upon this thesis to identify critical pain points in their applications, create an evidence-based prioritization framework based on the evolution of discovered issues, and employ focused recommendation strategies. In short, this thesis offers the means to move from a reactive problem-solving approach, to proactive decision-making initiatives.

Contents

1 Introduction

1.1 Background

Over the past decade, e-Government initiatives have significantly transformed public service delivery, increasingly relying on digital channels such as mobile applications, Web portals, and automated chatbots to streamline interactions between citizens and government institutions [63]. Within this global shift, the Netherlands has consistently positioned itself as a leader in digital governance. According to the United Nations E-Government Development Index (EGDI) from 2022, the Netherlands received a "Very High" (VH) rating across all key sub-indices—Online Service Index (OSI), Human Capital Index (HCI), and Telecommunication Infrastructure Index (TII)—achieving an overall EGDI score of 0.9384 and ranking among the top 10 countries worldwide [63]. These rankings underscore the country's strong commitment to digital transformation and highlight the importance of its digital applications as primary interfaces for delivering public services.

In this context, the digital ecosystem of Dutch government services continues to expand, including platforms for identity verification, official digital mailboxes, travel alerts, and secure document sharing. As highlighted by [52] and [24], the design of such public service applications must be citizen-centric. However, as these applications grow in both scope and complexity, maintaining high levels of user satisfaction presents an increasing challenge.

Optimizing user satisfaction is therefore not only a technological priority but also a core element of effective public service delivery. User satisfaction—reflecting what citizens like, think, feel, or wish to change about a digital service—can be quantitatively and qualitatively assessed through user feedback [35]. However, the sheer volume and unstructured nature of this feedback, often in the form of free text comments, make manual analysis impractical. This situation accentuates the need for advanced, automated models capable of extracting meaningful insights from large-scale user feedback, especially in highly digitalized contexts such as the Netherlands.

To effectively analyze such vast and diverse user-generated content, the field of Natural Language Processing (NLP) has emerged as a key facilitator. NLP provides computational techniques for automatically interpreting and deriving insights from human language, making it particularly well suited to extract structured patterns from unstructured feedback [27], [47]. In both academic and industry domains, NLP methods have been widely adopted to evaluate user satisfaction and optimize digital service delivery [8], [27]. Examples of these tasks are sentiment analysis, topic modeling, and issue classification.

To support such tasks, various modeling techniques have been developed to uncover patterns and structure within textual data. Traditional topic modeling methods such as Latent Dirichlet Allocation (LDA) have long served as foundational tools for identifying thematic patterns in large text corpora [4]. However, the advent of transformer-based architectures [64] has fundamentally reshaped the landscape of automated feedback analysis. The transformer model revolutionized NLP by introducing self-attenion, enabling superior capture of long-range dependencies and contextual relationships in text [64]. This architectural breakthrough paved the way for more complex language technologies, namely Large Language Models (LLMs). These models have further transformed feedback analysis through their ability to extract nuanced semantic insights and infer implicit user concerns from unstructured data. For public service agencies, this capability represents a paradigm shift: LLMs enable the development of more responsive, citizen-centric applications by translating raw data into actionable intelligence while preserving contextual integrity [5].

1.2 Relevant Theories on User Satisfaction and Digital Service Use

Understanding how users feel about government apps and why they report problems also involves looking at well-known theories from psychology and consumer research. While this thesis centers on adapting language technologies to interpret user experiences in digital public services, its methodological foundation can be rooted in these theoretical perspectives.

One important theory is the Expectation-Confirmation Theory (ECT), which explains satisfaction as the result of alignment between user expectations and actual service performance [42]. In posits that satisfaction is confirmed when the user's expectation and the performance of the service align [18]. This reaction is made stronger by the negativity bias, which means negative experiences often have a bigger impact than positive ones [3].

The Technology Acceptance Model (TAM) also helps explain why people adopt or abandon digital tools. It highlights ease of use and usefulness as key reasons people choose to use a technology [15], [9]. Thus, if a government app is difficult to navigate or doesn't seem helpful, people may stop using it.

Beyond individual-level satisfaction, broader frameworks like the Uses and Gratifications Theory (UGT) help explain user engagement and expectations. This theory categorizes user motivations into diversion, social utility, personal identity, and surveillance [28], which—when not fulfilled—can lead to dissatisfaction in task-oriented government services. Lastly, the SERVQUAL model outlines five dimensions of perceived service quality: reliability, responsiveness, assurance, empathy, and tangibles [44]. These dimensions offer a structured way to interpret user complaints not as isolated frustrations, but as patterns of unmet service and psychological expectations.

These theories clarify why users report issues, how dissatisfaction emerges, and what drives feedback. They show that negative experiences are more impactful (negativity bias), satisfaction stems from expectation alignment (ECT), and ease of use promotes engagement (TAM). They also explain how users seek to fulfill specific needs (UGT) and why uncertainty in service delivery causes frustration. This theoretical perspective allows the analysis to move beyond patterns and uncover the behavioral dynamics behind user satisfaction in public sector apps.

1.3 Research Problem

While user-generated feedback such as app store reviews offers rich, real-time insights into usability, functionality, and overall experience, its unstructured, high-volume, and multilingual nature renders manual analysis impractical and traditional NLP tools insufficient. Traditional classifiers and statistical models like LDA often require extensive human intervention and fail to capture the true-nature of issues expressed by users. This makes it difficult to implement a fully autonomous framework for analyzing and prioritizing public concerns. With LLMs' emergence, new opportunities have arisen for feedback analysis. However, their application in real-world government platforms remains largely unexplored. This research tackles the challenge of systematically levarging NLP systems—particularly LLMs—to extract, classify, assess, and predict issues in user reviews of government applications, offering a more responsive and citizen-aligned approach to public service improvement.

1.4 Research Gap

Despite growing interest in applying NLP to user feedback, several critical gaps persist in the current body of research within the NLP field, as will be further discussed in the next chapter (see Section 2).

First, issue extraction and topic modeling approaches frequently rely on predefined taxonomies or supervised techniques (e.g., [13], [22], [34], [65], [66]), limiting models' capacity to autonomously detect new or evolving issues, particularly in public sector contexts where concerns may shift over time. There is a notable lack of research investigating how LLM-based unsupervised methods can be used to extract, classify, and assess user feedback without human-labeled data or predefined categories.

Second, while traditional topic modeling techniques such as LDA are commonly used, comparative studies evaluating them against modern LLMs often lack standardized evaluation metrics (e.g., [36], [45], [65]). In addition, multi-label classification—essential for capturing the complexity of user feedback where multiple issues may co-occur in a single review— is underrepresented, with many studies simplifying review classification to single-label outputs (e.g., [1], [11]).

Third, although forecasting user sentiment is gaining popularity, existing work typically focuses on structured numerical data (e.g., [41], [57], [59]), with little attention paid to content-level forecasting. Furthermore, the literature lacks an integrated, end-to-end evaluation framework that encompasses issue extraction, classification confidence, star rating impact, temporal dynamics, and forward-looking recommendations.

This study addresses these gaps by proposing a scalable, unsupervised pipeline that leverages LLMs across the entire feedback analysis lifecycle, supported by a comprehensive and reproducible evaluation methodology.

1.5 Research Objective

This research aims to develop and evaluate an LLM-powered pipeline for extracting, classifying, evaluating, and forecasting user issues within the context of four Dutch government applications, namely: KopieID, Reisapp, MijnOverheid, and DigiD. The pipeline is designed to operate without predefined labels or supervised training data, enabling flexible and scalable insights across four core tasks:

1. **Issue Extraction**, to autonomously identify recurring topics and compare the coherence and granularity of LLM-derived issues against traditional models such as LDA;
2. **Review Classification**, aimed at performing multi-label classification of user reviews based on the previously identified issues;
3. **Issue-Star Rating Assessment**, to evaluate the relationship between identified issues and user satisfaction using Cumulative Link Models (CLMs) and temporal analysis;
4. **Forecasting**, to predict emerging issues and suggest improvements using LLM-generated insights based on past review data.

In line with these objectives, this research propose a structured framework that offers the Dutch government the means to shift user feedback analysis from a reactive, sentiment-driven process to a dynamic, forward-looking approach supporting continuous improvement of digital public services.

1.6 Research Questions

To address the identified research gaps and study objectives, this thesis introduces research questions that examine how modern NLP techniques, particularly LLMs, can extract, classify, and assess issues in user feedback from public service apps, and whether these language technologies can anticipate emerging concerns. Each main question is further explored through a set of sub-questions.

RQ1

How do LLMs extract and classify issues from user generated content compared to traditional methods?

1. How do LLMs compare to LDA in terms of coherence scores when extracting specific issues from user reviews? To what extent do the sets of identified issues overlap between the two methods?
2. How do the classification confidence scores differ between LLMs and LDA when categorizing app reviews?
3. To what extent do different LLMs agree on the categorization of issues within user reviews as measured by agreement metrics such as Krippendorf's alpha?

The first research question (RQ1) explores the capabilities of LLMs to autonomously identify and categorize issues in user feedback without relying on predefined taxonomies or labeled data. It contrasts these capabilities with those of traditional models like LDA, evaluating performance through coherence, classification confidence, and the overlap or divergence between LLMs and LDA. It relates to negativity bias [3] and SERVQUAL [44], which explain why negative service experiences often dominate user feedback and how they can be structured into quality dimensions. Moreover, this question and its sub-divisions link to the Uses and Gratifications Theory [28], which helps frame users' expectations and motivations for using digital public services.

RQ2

How do the extracted issues influence user satisfaction?

1. What relationships can be identified between specific issue types and user satisfaction metrics, and how do these relationships vary in significance across different application contexts?
2. How can temporal analysis reveal the evolving impact of different issues on user satisfaction over time?

The second research question (RQ2) explores how the extracted issues influence user satisfaction, measured through star ratings. It also considers how these relationships change over time and across application types. This question is grounded in the Expectation-Confirmation Theory [42], as it focuses on whether user expectations are met or violated, and how this affects satisfaction. The Technology Acceptance Model [15] also supports this inquiry by emphasizing that ease of use and usefulness are critical for continued app engagement.

RQ3

Can LLMs forecast future issues in user feedback and provide actionable insights to help businesses address emerging challenges and potentially improve user satisfaction?

1. How do LLM-based forecasts compare with insights derived from historical issue trends?
2. How can LLMs be leveraged to generate product improvement suggestions for businesses?

The third research question (RQ3) investigates the forecasting capabilities of these language technologies, assessing how well LLM-based predictions compare to historical trend analysis and how organizations can use these forecasts to proactively identify and address emerging user challenges. In addition, it explores the usefulness of the model-generated recommendations, evaluating their potential to steer practical enhancements that enhance user satisfaction and system reliability.

1.7 Thesis Structure

This thesis is organized into seven chapters. Chapter 2 reviews the relevant literature on sentiment analysis, topic modeling, classification (employing LDA, transformer-based methods, or LLMs), and textual forecasting, highlighting key gaps in current user-feedback methodologies. Chapter 3 outlines the analytical framework and experimental setup, detailing the methodology for the four core tasks—issue extraction, review classification, issue-star rating assessment, and forecasting—along with evaluation principles and key formulations. Chapter 4 introduces the four Dutch government applications studied (KopieID, Reisapp, MijnOverheid, and DigiD) and describes the preprocessing pipelines for handling multilingual feedback with LDA and LLMs. Chapter 5 presents empirical results for each task, including LLM–LDA comparisons in both issue extraction and classification tasks, issue-satisfaction analysis, and evaluation of forecasting and recommendation quality. Chapter 6 discusses these findings in relation to the research questions, and Chapter 7 concludes with key contributions, limitations, and future research directions.

2 Related Work

In the field of NLP, user feedback analysis has been widely explored using data from applications, movies, product reviews, and tweets to achieve various objectives. These studies primarily focus on interpreting and understanding user satisfaction, predicting the likelihood of purchasing similar or related products, and identifying key topics discussed in user reviews. In this chapter, an overview of related research is presented, organized into three thematic areas relevant to this study. Each area highlights its connection to this thesis, summarizes key findings from prior work, and outlines limitations that this study aims to address and build upon. Moreover, each section will conclude with a bar-marked paragraph summarizing its main takeaway, emphasizing the primary challenges associated with the specific task under discussion.

2.1 Sentiment Analysis in User Feedback Interpretation

Although the focus of this thesis is not on extracting users sentiment from reviews, the process of analyzing reviews to determine user positivity or negativity shares similar methodologies, models, and preprocessing steps with those employed in this study. Additionally, the effets of the identified issues on star ratings, used here as a proxy for user sentiment, were examined. Accordingly, a discussion of sentiment analysis was considered a valuable addition to this thesis.

Sentiment analysis, often referred to as opinion mining [43], involves identifying and interpreting the emotional tone or attitude conveyed in a text. The analysis can be a more restricted one focusing on distinguishing only between negative, positive or neutral texts [16], [31], or a more nuanced one, focusing in addition on emotions such as joy, frustration, rage, or sadness [36].

The authors in [67] focused on fine-tuning transformer-based models, such as BERT, RoBERTa, and DistilBERT, on the IMDB movie review dataset to assess their effectiveness in sentiment classification. The study emphasized the significant impact of the maximum sequence length parameter on model performance and detailed preprocessing steps, including lemmatization, stemming, and the removal of URLs, punctuation, and stopwords. However, it found that excessive cleaning could hinder the models' ability to capture contextual nuances essential for sentiment analysis. While the paper provides valuable insights, including an ablation study on the `max-len` parameter for transformer models, it lacks dataset diversity, limiting the generalizability of its findings. Nonetheless, it offers an informative exploration of transformer-based models in sentiment classification tasks.

An earlier study [31] introduced the BERT-BiGRU-Softmax model for sentiment analysis of e-commerce product reviews and performed a comparative analysis against traditional models, including standard BERT, demonstrating that the new model outperformed its counterparts. However, the study has several limitations. First, the analysis was conducted on reviews in a single language, presumably Chinese, without any mention of a translation process, limiting its applicability to multilingual contexts. Second, the study relied on predefined dimensions (e.g., price, quality, service) for sentiment analysis, meaning the main topics were predetermined. This constraint hinders the model's ability to autonomously discover new aspects or issues that might impact users' satisfaction.

LLMs have also been explored for sentiment classification tasks. In [45], models such as Falcon 7B, considered state-of-the-art, and GPT-2 were used for supervised learning on a labeled dataset of TripAdvisor consumer reviews. The dataset was annotated by English-proficient students with positive or negative labels and underwent a preprocessing phase before training. Another study ([34]) examined more recent ChatGPT versions, including ChatGPT-3.5, GPT-3.5-turbo, and GPT-4.0, to evaluate their performance in sentiment analysis. However, this study was limited to only two product reviews (a budget smartphone and a bike), and no attempts were made to explore unsupervised learning approaches.

On a more theoretical level, the authors of [30] introduced the integration of causal reasoning with LLMs, focusing on addressing questions like “why is the user not satisfied.” The study demonstrates that LLMs significantly enhance the accuracy of sentiment analysis in e-commerce by effectively processing and interpreting unstructured customer feedback. Furthermore, combining LLMs with causal reasoning techniques provides a deeper insight into the underlying factors influencing customer sentiments. However, the paper falls short in practical applications, focusing primarily on benchmarking metrics rather than exploring real-world implementations.

Moreover, in [21], a theoretical overview of the sentiment analysis task was presented, tracing its evolution from traditional rule-based approaches to modern large language model-based techniques. The study highlights key challenges in sentiment analysis, such as handling bilingual texts, and examines the impact of LLMs on the field. However, it focuses solely on theoretical discussions without presenting any experimental results, models, or datasets. Additionally, while acknowledging the complexities of bilingual sentiment analysis, the study focuses solely on mBERT, neglecting alternative approaches such as translation methods or multilingual LLMs.

> The sentiment analysis paradigm, as highlighted in the literature review, faces several key challenges: handling bilingual or multilingual content in reviews, reliance on sentiment analysis based on predefined topics within the text, the constant need for annotated datasets, and the predominance of supervised methods. This reliance on supervised approaches constrains the analysis to either clean, labeled datasets or requires significant human intervention, limiting scalability and adaptability in more complex or dynamic contexts.

2.2 Topic Modelling & Classification Techniques for Feedback Analysis

In [34], the aforementioned LLMs are used not only to assess user satisfaction but also to perform topic assignment for each review. However, the study does not address topic extraction and instead relies on pre-defined topics to create prompts, limiting the models' ability to autonomously detect and generate topics from the data. Similarly, the authors from [36] employ the LDA algorithm from the `gensim` [17] library's native LdaModel, which was also considered as a benchmark in this paper. However, the study does not evaluate its performance against alternative NLP methods. Furthermore, it fails to explicitly detail the evaluation metrics used to validate either sentiment analysis or topic modelling tasks, limiting the reproducibility and interpretability of its findings.

The study by Praveen et al. [45], also relevant to this section, evaluated the performance of state-of-the-art models, specifically the LLM Falcon 7B and the transformer-based model BERT, in both sentiment analysis and topic modeling. One of their goals was to compare LLM-based topic modeling with the traditional LDA approach, using a methodology closely aligned with this thesis. However, the study did not include newer models like LLaMA 3 or GPT (limitation acknowledged by the authors too), lacked evaluation metrics such as coherence scores, and was limited to a single dataset, restricting the generalizability of its findings.

The topic modeling task is often closely linked to text classification, and this combination has been explored in numerous studies over the years. For instance, in [1], the authors combined topic modeling using LDA with multi-class classification. LDA was used to extract dominant topics from Amazon baby product reviews, and the Machine Learning (ML) classifiers categorized these reviews into predefined product categories based on the extracted topics and features. Their classification results underscored the challenges of achieving high performance with traditional ML methods, such as Support Vector Machine (SVM), Logistic Regression, and Naïve Bayes, on product review datasets. However, one key limitation of their approach was the lack of consideration for multi-label classification, where a single review could be linked to multiple topics. Instead, each review was assigned only one dominant topic, overlooking the possibility of co-occurring or interconnected topics. This approach restricted the depth of the analysis, as it failed to capture the nuanced relationships between topics within the reviews. A similar issue was identified in [11], where the AR-Miner framework was used to classify app reviews as "informative" or "non-informative". Despite its utility, the framework also assigned reviews addressing multiple topics to a single topic, failing to account for the complexity of multi-topic app reviews.

An earlier study from 2014 addressed the challenge of multi-label classification, where reviews can belong to multiple topics, by employing the "one-vs-all" strategy, training a separate classifier for each category [13]. While effective, this approach is computationally expensive. The primary contribution of the paper is a framework that integrates user reviews with mobile app code analysis to improve release planning by aligning user expectations with development priorities. However, the topic modeling process was constrained by manually defined taxonomy delimitations, rather than being automatically extracted. This raises questions about the framework's generalizability, even though the study used data from 39 open-source apps from Google Play Store.

Additionally, multi-label classification was employed in [22], alongside a novel segmentation algorithm derived from LDA, called TopicDiff-LDA. This approach improved the annotation of customer reviews by segmenting multi-topic documents into semantically coherent units and fine-tuning an LDA model based on these segments. Although this method successfully addressed the issue of missing relationships between topics, the study relied solely on traditional classifiers without leveraging more advanced models such as transformers or LLMs. Another limitation was the manual generation of topic labels based on the top terms extracted, introducing a dependency on human interpretation and reducing adaptability to new or unforeseen topics.

More recent studies, such as [65] and [66], have adopted LLMs for classification tasks instead of traditional ML models on e-commerce datasets or on reviews from Google Play, demonstrating that LLMs outperform traditional methods in such applications. However, both papers relied on labeled datasets, with [66] using manual annotations. The same paper highlighted that LLMs like ChatGPT can perform bilingual app review mining without requiring additional fine-tuning, leveraging their pre-trained capabilities for zero-shot or few-shot learning.

The connection between sentiment analysis, topic modeling, and topic classification was explored in [19], where the Structural Topic Modeling (STM) algorithm was used to identify and categorize topics in sentiment-labeled hotel reviews from Booking.com. These reviews were divided into two categories based on hotel ratings (2–3 stars and 4–5 stars), and the XGBoost classifier was subsequently employed to classify each review. However, using sentiment analysis as a proxy for revisit intentions introduces potential inaccuracies, as positive sentiment does not always correlate with the intention to revisit. Additionally, the study overlooks temporal patterns, which could provide valuable insights into how revisit intentions evolve over time.

> The following challenges have been consistently highlighted in the literature regarding topic assignment, modeling, and classification: topics are often predefined through human intervention and explicitly used in prompts, limiting the models' ability to autonomously identify new topics. The evaluation of topic modeling tasks typically relies on comparisons with traditional models, without standardized metrics for assessment. Additionally, multi-label classification is frequently overlooked, leading to the neglect of topic co-occurrence and interconnectivity. Even when multi-label classification is considered, it is often implemented using the "one-vs-all" strategy, which is both computational and time-intensive.

2.3 Forecasting Potential of LLMs: User Feedback-Based Prediction

According to the analysis presented in [57], which reviewed multiple studies, LLMs have shown remarkable potential in parsing and analyzing large datasets, such as the Amazon Review dataset, to identify patterns, predict future trends, and detect anomalies across various domains. Among the LLMs discussed, GPT-3, GPT-3.5, GPT-4, LLaMA2-7b, LLaMA2-13b, and LLaMA2-70b were specifically employed for forecasting tasks.

Prediction tasks have been successfully tackled by LLMs, particularly on time series data [59] and grid-world mazes [41]. The focus of this thesis aligns more closely with the former, as it shares greater similarity with one of the tasks presented in this paper. The experimental study in [59] explores the zero-shot application of LLMs for time series forecasting without fine-tuning, a methodology similar to the developed in this thesis, though applied to text datasets rather than numerical ones. The study demonstrates that LLMs excel at forecasting time series with clear patterns and trends but face challenges with datasets lacking periodicity or containing multiple overlapping patterns. Additionally, the models are particularly sensitive to the most recent segments of input sequences, highlighting potential limitations in handling complex or irregular time series.

Another study [55] focused on event prediction, where LLMs, guided by a few expert-annotated examples, suggest possible causes for proposed events, providing insights into potential future purchases. The authors developed the LAMP framework, which integrated various LLMs, including GPT-3-davinci, GPT-3.5-turbo, and LLaMA-2-chat, and tested it on three different datasets. While this study shares some similarities with the present work, its focus is on predicting the next purchase rather than identifying issues users have experienced with products. Furthermore, the framework's performance heavily relies on high-quality annotated examples, which poses scalability challenges in domains with limited expert annotations, such as app reviews.

> The main takeaways regarding the forecasting paradigm using LLMs are as follows: the focus has predominantly been on analyzing numerical datasets, which present significant challenges when lacking periodicity or containing multiple overlapping patterns. Conversely, when applied to review-based datasets, the emphasis is primarily on predicting the next purchase rather than identifying issues mentioned in the content or forecasting potential problems with the product in the future.

3 Methodology

3.1 Overview

This study focuses on four distinct yet interconnected components: (1) extracting issues or topics from user reviews of four government applications, (2) classifying each review according to the previously identified issues, (3) analyzing the relationship between the extracted and classified issues and the corresponding star ratings, and (4) forecasting potential future issues by leveraging historical reviews and their associated issues from before 2024.

The central element of this study is the use of NLP methods, in particular LLMs. LLMs are advanced machine learning systems trained on vast amounts of textual data to understand and generate human-like language. These models, often based on transformer architectures and considered a class of foundational models [25], are capable of performing a wide range of natural language processing tasks, such as summarization, translation, question answering, and text classification [54]. Their ability to generalize, follow instructions, handle unstructured input, and generate contextually relevant outputs [68] makes them particularly suitable for analyzing feedback data from public sector applications. The performance and interpretability of these models are evaluated and compared throughout the course of this research.

For the first task, five different LLMs, both proprietary and open-source have been used: GPT-4o-mini, Claude-3.5-Sonnet-202411022, Gemini-1.5-Pro, Gemini-2.0-Flash, and Mistral-Large-2411. These models were selected based on their Chatbot Arena LLM Leaderboard scores as observed in February 2025. At that time, the specified versions of GPT, Gemini, Claude, and Mistral had the Arena Scores for the Hard Prompts (English) and Instruction Following categories, as shown in Table 1 [56]. However, cost considerations also played a role in model selection. As a result, the second task was restricted to four models, excluding Claude 3.5 due to its high costs and limited API request quota. Moreover, for the third and fourth tasks only the best-performing LLM was used. The criteria for determining the *best* model will be elaborated further in this chapter. Additionally, LDA [4] was also employed for the first two tasks to benchmark the performance of LLMs in extracting and classifying issues from reviews.

Model	**Arena Scores**	
	Hard Prompts (English)	**Instruction Following**
gpt-4o-mini	1258	1266
claude-3-5-sonnet-20241022	1293	1303
gemini-1.5-pro	1243	1262
gemini-2.0-flash	1347	1352
mistral-large-2411	1245	1256

Table 1: Arena scores for the used models in February 2025 across two tasks: Hard Prompts (English) and Instruction Following. These values represent model performance in head-to-head comparisons based on human preferences, evaluated using the Bradley-Terry (BT) model [12]. The resulting scores lie on a continuous, relative scale with no fixed maximum or minimum—higher scores indicate stronger overall performance.

This study used multiple datasets of various sizes to support the analysis. Each dataset consists of user reviews from one of the four government applications: KopieID, Reisapp, MijnOverheid, and DigiD. These datasets contain star ratings, timestamps, and textual feedback from users, covering multiple years. The datasets were preprocessed to remove duplicate entries, filter out irrelevant content, and translate all reviews into English to ensure consistency before analysis with the selected LLMs. Further details on the datasets, their collection process from Google Play, and preprocessing steps will be presented in Section 4.

Figure 1 provides an overview of the methodology used in this thesis, outlining the entire pipeline. Building on this framework, the next sections will present the experimental setup in detail, along with the evaluation metrics used.

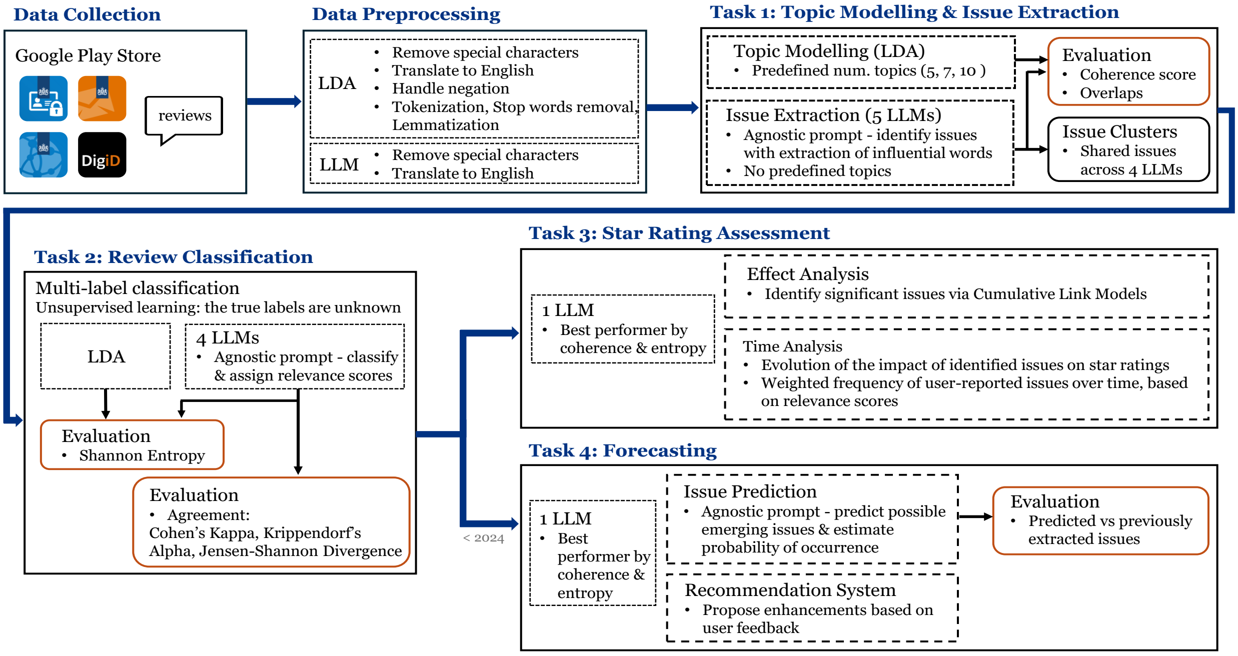

Data Collection
Google Play Store
reviews
DigiD
Data Preprocessing
LDA
Remove special characters
Translate to English
Handle negation
Tokenization, Stop words removal, Lemmatization
LLM
Remove special characters
Translate to English
Task 1: Topic Modelling & Issue Extraction
Topic Modelling (LDA)
Predefined num. topics (5, 7, 10)
Issue Extraction (5 LLMs)
Agnostic prompt - identify issues with extraction of influential words
No predefined topics
Evaluation
Coherence score
Overlaps
Issue Clusters
Shared issues across 4 LLMs
Task 2: Review Classification
Multi-label classification
Unsupervised learning: the true labels are unknown
LDA
4 LLMs
Agnostic prompt - classify & assign relevance scores
Evaluation
Shannon Entropy
Evaluation
Agreement:
Cohen's Kappa, Krippendorf's Alpha, Jensen-Shannon Divergence
Task 3: Star Rating Assessment
1 LLM
Best performer by coherence & entropy
Effect Analysis
Identify significant issues via Cumulative Link Models
Time Analysis
Evolution of the impact of identified issues on star ratings
Weighted frequency of user-reported issues over time, based on relevance scores
< 2024
Task 4: Forecasting
1 LLM
Best performer by coherence & entropy
Issue Prediction
Agnostic prompt - predict possible emerging issues & estimate probability of occurrence
Recommendation System
Propose enhancements based on user feedback
Evaluation
Predicted vs previously extracted issues

Figure 1: Methodology Overview

3.2 Experimental Setup

This section describes the experimental setup, covering each stage of the analysis: issue extraction using LDA and LLMs, classification of user reviews, evaluation of issue impact on star ratings, and forecasting of future issues.

3.2.1 Issue Extraction & Evaluation Methodology

For the issue extraction process, two complementary approaches were used: traditional topic modeling using LDA and LLMs. This dual approach enabled both a benchmark comparison and a richer understanding of the thematic structures within the user reviews.

LDA was applied to each dataset (KopieID, Reisapp, MijnOverheid, and DigiD) using the `gensim` Python package. As a widely used probabilistic topic modeling technique, LDA was selected for its ability to uncover latent thematic patterns in textual data. The algorithm was executed with 5, 7, and 10 topics to assess how different topic granularity levels impact the extracted representations. Establishing LDA as a baseline provided a clear reference point for evaluating the issue extraction capabilities of LLMs. It helped determine how closely LLM-generated issues align with traditional topic modeling and whether they provide better context and detail.

For LLM-based extraction, each model received an agnostic prompt without predefined constraints regarding the type or number of issues, allowing it to freely identify relevant and recurring problems based on its understanding of the text. The prompt also asked for a list of influential words per issue, offering insight into the model's reasoning and supporting later evaluations, such as coherence scoring and embedding-based similarity. This setup enabled a structured, unbiased comparison between LLM-based and traditional methods.

The effectiveness of each LDA configuration was, thus, evaluated using coherence scores, which measure the semantic consistency of topics by analyzing word co-occurrence patterns. Similarly, to assess the quality of the issues extracted by the LLMs, the influential words generated by the models were utilized as the foundation for evaluation, providing insights into the coherence and relevance of the identified topics. These words were transformed into word embeddings using the Term Frequency-Inverse Document Frequency (TF-IDF) method, allowing for a quantitative analysis of the semantic relationships between different extracted issues.

Further the coherence scores were recomputed to determine whether the LLMs identified well-structured and meaningful topics or merely produced collections of unrelated terms. This step also allowed a direct performance comparison between the LLM-generated issues and those extracted using LDA. To ensure a fair comparison between the LLM-extracted issues and those derived by the LDA when computing the coherence scores, a dictionary based on the preprocessed reviews from the LDA step was used. The LDA-preprocessed reviews underwent extensive text cleaning as mentioned in Chapter 4, making them more suitable for coherence analysis. Using this cleaned dataset also helped prevent bias in the evaluation process by ensuring that the same linguistic foundation was used to compare the extracted issues.

Additionally, a heatmap was generated to visualize the similarity between the LLM-extracted issues and the LDA topics across all models and datasets. By comparing the TF-IDF embeddings of the extracted issues, the cosine similarity between each LLM's results and LDA's structured topic modeling approach was computed. This visualization provided an intuitive way to determine whether LLMs captured the same underlying themes as LDA or if they introduced additional nuances.

Finally, to further enrich the analysis, the alignment between the outputs of different LLMs was assessed by clustering the extracted issues. This approach helped identify shared topics across models, uncover their underlying categories, and explore how they group into broader themes.

3.2.2 Review Classification & Evaluation Methodology

Once the issues were extracted, each LLM was tasked with classifying the reviews according to these issues. Again, an agnostic prompt was used, ensuring that the models categorized the reviews freely without predefined constraints. Given the complex nature of user reviews, where multiple concerns are often expressed within a single entry, a multi-label classification approach was employed. This method enabled a review to be associated with multiple issues simultaneously, reflecting the reality that user feedback is rarely confined to a single theme.

To quantify the relationship between a review and the extracted issues, the LLMs were instructed to assign a percentage relevance score to each issue, indicating the degree to which a given problem was present in the review. However, multi-label classification often results in relevance scores that sum to more than 1, as individual labels are assigned independently rather than in a mutually exclusive manner. Such behavior is a known characteristic of multi-label learning and does not indicate an issue with classification quality [62]. However, to ensure consistency in output structure and facilitate direct comparison among models, a normalization technique was applied to standardize the classification results. Normalization ensured that the classification results remained comparable across different models (including LDA), preventing excessive weighting of certain issues and allowing for a more interpretable comparison between LLM outputs.

Since the classification was conducted in an *unsupervised* manner, without predefined labels or ground truth references, a diverse set of evaluation metrics was employed to systematically assess the performance of the LLMs from different angles.

The Shannon Entropy was computed to measure the confidence level of each LLM in performing multi-label classification, offering insights into the stability and reliability of the model outputs. Entropy is a widely used metric in information theory that quantifies the degree of uncertainty in a probability distribution [53]. In this context, higher entropy values indicated higher uncertainty, suggesting that the LLMs distributed relevance scores more evenly across multiple issues, possibly reflecting ambiguity in classification. Conversely, lower entropy values suggested that the LLMs assigned issues with greater certainty, concentrating their predictions on a smaller set of dominant topics.

Finally, an additional step was taken to compare the classification performance of the LLMs relative to one another and to evaluate the consistency and alignment between LLMs in multi-label issue classification tasks. To this end, three agreement metrics—Jensen-Shannon Divergence (JSD), Cohen's Kappa, and Krippendorf's Alpha—were employed to assess the similarity between topic distributions produced by each model. By combining these three metrics, the evaluation captures both hard agreement (presence/absence of issues) and soft agreement (magnitude of predicted relevance), while also adjusting for chance agreement and accommodating models with varying outputs.

3.2.3 Issue-Star Rating Assessment & Evaluation Methodology

This analysis integrates two complementary approaches to evaluate the relationships between issues and user satisfaction, as measured through app star ratings. Specifically, it combines a Cumulative Link Model analysis for precise estimation of issue impact, and a temporal analysis to track how these impacts evolved over time.

The CLM, a type of regression model specifically developed for analyzing ordinal-scale data, estimate the probability that an observation falls into or below a certain category by modeling cumulative probabilities through a link function [23]. In this study, the star ratings (ranging from 1 to 5) act as the ordinal dependent variable, while the identified issues and their relevance scores serve as explanatory variables. CLMs are particularly appropriate for this scenario, as they estimate how the presence of specific issues influences the probability of a user assigning a higher or lower star rating.

To ensure methodological rigor, the input data for this analysis consisted of the issue classifications and relevance scores produced by the best-performing model, as identified during the issue extraction and classification stages.

Observation 1:

The selection of the best-performing model was based on three key criteria. First, the model needed to demonstrate a high coherence score for the extracted issues, ensuring that the topics were semantically meaningful and well-structured. Second, the Shannon entropy distribution of its classifications was evaluated, with preference given to models exhibiting lower entropy values, indicating greater classification confidence. Together, these two factors: coherence and confidence, formed the core of the selection process. In addition to these primary criteria, further considerations were applied to refine the selection. The model's reliability was assessed by examining hallucinations, defined as instances where the model produced erroneous or irrelevant outputs. Moreover, adherence to the expected structured output format (JSON) was evaluated, as maintaining consistency in formatting was essential for ensuring the accuracy and interpretability of downstream analyses.

Moreover, following the methodology outlined in [2], multiple link functions were tested to determine the most suitable model structure for each application dataset. Specifically, five link functions were evaluated: logit, probit, complementary log-log (cloglog), log-log (loglog), and cauchit.

The Akaike Information Criterion (AIC) was calculated for each model, and the link function yielding the lowest AIC value was selected.

By quantifying how each user-reported issue affects star ratings, organizations can identify the problems that exert the greatest negative influence on perceived app quality. This allows them to identify not only which problems are most frequent, but also which have the greatest potential to damage user perception. Ultimately, this enables a proactive approach to quality management, focusing on mitigating high-impact issues before they escalate into widespread dissatisfaction.

To complement the regression analysis and provide a dynamic view of user sentiment, a time-based impact analysis was also conducted. As app ratings serve as key indicators of user satisfaction, their fluctuations in response to identified issues were analyzed to assess which problems had the greatest impact on user perception.

For consistency and reliability, the same LLM-extracted and classified issues were used in the time analysis, ensuring alignment with the CLM stage. Thus, a quantitative impact assessment was performed by correlating the frequency of each issue with observed changes in app ratings over time. This approach enabled the distinction between short-term spikes in negative ratings and long-term trends.

3.2.4 Forecasting & Evaluation Methodology

With the issue extraction, classification, and issue-star rating assessment stages completed, and the best-performing model identified, the next step of the analysis focused on forecasting potential future issues. Using this selected model, the task was to predict possible emerging issues, while also estimating the probability of occurrence for each forecasted issue.

By leveraging the patterns identified in historical user reviews, the model aimed to extrapolate potential concerns that users might express in subsequent months or years. This predictive capability was particularly valuable for government applications, as it allowed for proactively addressing concerns before they escalated into widespread problems. Beyond forecasting, the LLM also generated recommendations for mitigation strategies. These recommendations provided a data-driven foundation for decision-making, enabling government to improve their applications based on predicted user concerns rather than just retrospective analysis.

To validate the forecasted process, the predicted issues were systematically compared with actual issues that emerged in 2024 and beyond. This evaluation process involved several key steps. First, the forecasted issues were vectorized into numerical embeddings using the *all-MiniLM-L6-v2* sentence transformer, a lightweight yet powerful model optimized for generating high-quality sentence embeddings [48], [50]. Similarly, actual issues that surfaced after 2023 - extracted using the same issue extraction methodology - were also transformed into embeddings. Once both sets of issues were represented in this way, their cosine similarity was computed to measure how closely the forecasted issues aligned with real-world concerns that users had later reported.

This evaluation approach was critical in assessing the practical utility of LLM-based forecasting. The forecasted and actual issues were compared to assess whether the model's predictions were speculative or aligned with observable trends. This analysis also revealed which types of issues were most predictable and which were harder to anticipate due to shifting user expectations, technological changes, or external factors.

3.3 Quantitative Evaluation Criteria

The methodology employed in this study is based on unsupervised learning, meaning that the datasets lack predefined true labels, making evaluation a particularly challenging task. Unlike supervised approaches, where model performance can be assessed against a ground truth, unsupervised methods require alternative strategies to measure their effectiveness. In the previous sections, the key concepts and criteria used to assess the LLMs' performance were introduced. However, the formal mathematical definitions of these evaluation metrics have been reserved for this section, where their theoretical foundations and practical relevance will be analyzed in more depth.

3.3.1 Coherence Score

The coherence score is a fundamental evaluation metric in unsupervised learning methodologies such as text mining and topic modeling [49], used to assess the semantic consistency of words grouped within a topic. In the context of this study, it quantifies the ability of LLMs and LDA to extract meaningful and interpretable issues from user reviews. A higher coherence score suggests that the words associated with a topic or issue are more semantically related, making the extracted topics more interpretable and contextually meaningful.

For this analysis, the C_V coherence score, as implemented in the `gensim` Python package, was employed. According to the framework proposed in [49], the C_V metric is defined by the tuple $(S_{set}^{one}, P_{sw(110)}, \tilde{m}_{cos(nlr,1)}, \sigma_a)$, where each element specifies a core component of the coherence calculation. S_{set}^{one} denotes a segmentation strategy where all word pairs are drawn from a single set of topic words. $P_{sw(110)}$ represents the probability estimation method, using a sliding window of size 110 to compute co-occurrence statistics. $\tilde{m}_{cos(nlr,1)}$ defines the confirmation measure, applying indirect cosine similarity to normalized context vectors built from co-occurrence information. Finally, σ_a specifies the aggregation function, which takes the arithmetic mean of all confirmation scores to produce the final coherence value.

Additionally, a more detailed explanation of the metric, including its underlying mathematical formulations used at each stage of its computation will be provided in the following lines. As introduced by [58], the C_V metric is computed through a sequence of four structured steps, which support the tuple definition presented above.

1. **Segmentation Phase**
   In this first step, known as the segmentation phase, the top-N most relevant words from a topic – denoted by $W = \{w_1, \ldots, w_N\}$ – are used to generate word pairs for further analysis. The segmentation is based on the one-set strategy, where each word $w_i \in W$ is evaluated in relation to the entire word set W. Formally, for each w_i, a pair is created as (W', W^*), where $W' = \{w_i\}$ is a singleton set containing the target word, and, in this case $W^* = W$ is the complete set of topic words. Finally, the set of all such segmentations is defined as: $S = \{(W', W^*) \mid W' = \{w_i\}, W^* = W, w_i \in W\}$. This approach corresponds to the segmentation notation S_{set}^{one} , indicating that each word is individually compared to the entire topic word set.

2. **Probability Estimation**
   In this stage, the co-occurrence probabilities of individual words w_i and word pairs (w_i, w_j) are estimated. These probabilities are computed as the number of documents in which the word w_i, or the pair (w_i, w_j) occurs, divided by the total number of documents. Unlike traditional document-level co-occurrence, the C_V metric introduces the concept of "virtual documents" [58]. This is achieved by applying a Boolean sliding window of fixed size $s = 110$ word tokens across the original documents. Each window is treated as a separate virtual document, enabling a finer-grained estimation of co-occurrence patterns that also considers the proximity of words within the local context. The final probabilities $p(w_i)$ and $p(w_i, w_j)$ are then computed by aggregating occurrences across all generated windows.

3. **Confirmation Measure**
   The third step involves calculating the confirmation measure. This metric quantifies the degree

to which the context represented by W^* supports or confirms the meaning of W'. In essence, it evaluates the semantic similarity between the two word sets relative to the overall context W.

To compute this similarity, both W' and W^* are transformed into context vectors based on Normalized Pointwise Mutual Information (NPMI) values. The NPMI quantifies how strongly two words are associated, normalized to account for word frequency biases [6]. The exact formulation of NPMI is provided in equation 1 and it is equivallnet to the confirmation of a single pair S_i [49]. The construction of the corresponding context vectors is detailed in equation 2. In these formulae, two additional parameters are introduced: ϵ, which prevents division-by-zero errors in the NPMI calculation, and γ, which applies greater weight to higher NPMI values, emphasizing stronger word associations.

$$NPMI(w_i, w_j) = \frac{\log \frac{P(w_i,w_j)+\epsilon}{P(w_i)\cdot P(w_j)}}{-\log(P(w_i, w_j) + \epsilon)} = \tilde{m}_{nlr}(S_i) \tag{1}$$

$$\vec{v}(W') = \left\{ \sum_{w_i \in W'} NPMI(w_i, w_j)^{\gamma} \right\}_{j=1,\dots,|W|} \tag{2}$$

Once the context vectors for W' and W^* are generated, the indirect confiramtion is computed using the cosine similarity between the two vectors (see equation 3). This measure captures the semantic closeness between the word sets in a high-dimensional space, serving as the core similarity function in the C_V coherence computation.

$$\phi_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2} = s_{cos}(\vec{u}, \vec{w}) \tag{3}$$

4. **Aggregation**

   In the final step of the C_V coherence computation, the individual confirmation measures calculated in the previous phase are aggregated into a single coherence score. This is done using the arithmetic mean, denoted as σ_a. Formally, the final coherence score is computed as: $C_V = \frac{1}{N} \sum_{i=1}^{N} \phi_{S_i}(\vec{u}, \vec{w})$.

Finally, the decision to use the C_V coherence metric in this study is supported by the empirical findings from [49], where various coherence measures were systematically evaluated. According to their experiments, the C_V metric was considered to be the best-performing one. Its ability to combine context-sensitive similarity (via cosine similarity) with normalized co-occurrence statistics (through NPMI and sliding windows) makes it particularly well-suited for evaluating the semantic interpretability of topics in unsupervised settings. All these reasons make the C_V an adequate primary metric for assessing the topic and issue coherence in this work.

3.3.2 Cosine Similarity

The mathematical definition of cosine similarity between two vectors $\vec{u}$ and $\vec{w}$ has already been introduced in equation 3. However, its interpretation and practical significance are addressed here in more detail.

Cosine similarity is a fundamental metric in vector space models and it is widely employed in natural language processing, text mining, and information retrieval to measure the semantic similarity between two vectors [51]. In this study, cosine similarity is used to compare issue vectors—constructed from the most relevant words extracted by various models—serving two main purposes: (1) to analyze the similarity between issues identified by different LLMs or LDA, and (2) to evaluate how closely the forecasted issues align with the actual issues found in later user feedback.

Unlike distance-based metrics, cosine similarity measures the orientation of vectors in high-dimensional space rather than their magnitude, making it particularly well-suited for comparing text-derived embeddings [61]. These embeddings can be generated using a variety of techniques, including word2vec, TF-IDF, or transformer-based models such as BERT. In this study, TF-IDF was employed for constructing word embeddings in the first two tasks (issue extraction and classification), while sentence embeddings based on the *all-MiniLM-L6-v2* transformer model were used for forecasting, as previously mentioned.

3.3.3 Term Frequency - Inverse Document Frequency

Term Frequency–Inverse Document Frequency is a widely used statistical technique for representing textual data in a vector space. It is a well-established approach in information retrieval systems [46] and is also commonly employed in text classification [14] and topic modelling tasks [1], [13].

TF-IDF balances two components: Term Frequency (TF), which measures how often a word appears in a document, and Inverse Document Frequency (IDF), which reduces the weight of commonly occurring words across the corpus [26]. The resulting value highlights words that are both frequent in a document and rare across documents, making them more meaningful for distinguishing content.

Mathematically, the TF-IDF weight of a term *t* in a document *d* can be written as in equation 4, where $TF(t, d)$ is the frequency of term t in document d, N is the total number of documents in the corpus, and $DF(t)$ is the number of documents containing the term t.

$$TF - IDF(t, d) = TF(t, d) \times \log\left(\frac{N}{DF(t)}\right) \tag{4}$$

3.3.4 Shannon Entropy

Shannon entropy is a fundamental concept in information theory used to quantify the uncertainty or unpredictability in a probability distribution [53].

In the context of this study, Shannon entropy is used as a metric to evaluate the confidence of the multi-label classification performed by the LLMs. Thus, when an LLM assigns relevance scores across multiple issues for a given review, a low entropy value indicates that the model was confident in its prediction, favoring one or few issues strongly, whereas a high entropy value suggests a more ambiguous or uncertain classification, with relevance scores distributed more evenly across issues.

The entropy H of a probability distribution $P = \{p_1, p_2, ..., p_n\}$, where each p_i represents the normalized relevance score assigned to an issue, is defined in equation 5.

$$H(P) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{5}$$

In the formula presented in 5, $p_i \log_2 p_i$ measures the information content of the outcome p_i, and the summation over all n issues yields the total uncertainty of the distribution. The maximum value of the Shannon entropy can be derived by considering a uniform probability distribution, where all n possible issues are equally likely, such that $p_i = \frac{1}{n}$ for all i. Substituting this into the entropy formula yields: $H(P) = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n$. Therefore, the maximum entropy increases logarithmically with the number of possible labels. In the context of this multi-label classification task, where each review can be associated with multiple issues, the number of potential labels exceeds two, and as a result, entropy values greater than 1 are expected and indicate higher uncertainty in distributions with many possible outcomes.

3.3.5 Jensen-Shannon Divergence

Jensen–Shannon Divergence is an information-theoretic metric used to measure the similarity (or the divergence) between two probability distributions. The JSD is a symmetric and smoothed version of the Kullback-Leibler divergence, measuring how close two probability distributions are while ensuring finite and bounded results [7]. However, the symmetry property of the JSD is only guaranteed when the two probability distributions being compared contain the same number of finite elements. In this study, each LLM may detect a different number of issues, resulting in distributions of unequal dimensionality. Consequently, the symmetry of the metric may not always hold in practice.

Given two discrete probability distributions P and Q, the Jensen–Shannon Divergence is defined in equation 6, where M is the average distribution ($M = \frac{1}{2}(P+Q)$) and $D_{\mathsf{KL}}(P \parallel M)$ and $D_{\mathsf{KL}}(Q \parallel M)$ are the Kullback–Leibler divergences between each original distribution and the average. Moreover, the Kullback–Leibler divergence is expressed in equation 7. Its values are in the range $[0, 1]$ when using the base-2 logarithm, where 0 indicates that the distributions are identical and values closer to 1 reflect increasing disimilarity between the distriutions.

$$JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M) \tag{6}$$

$$D_{\mathsf{KL}}(P \parallel Q) = \sum_i P(i) \log_2 \left(\frac{P(i)}{Q(i)} \right) \tag{7}$$

3.3.6 Cohen's Kappa

Cohen's Kappa is a statistical measure of inter-rater reliability, which quantifies the degree of agreement between two annotators classifying items into categories [60]. Unlike simple accuracy, Cohen's Kappa takes into account the agreement occurring by chance, induced by the possibility that some raters guess in the case of uncertainty [32], making it a more robust metric for evaluating classification consistency. In the context of this study, Cohen's Kappa is used to assess the degree of alignment between issue classifications generated by different LLMs.

It is important to note that Cohen's Kappa is originally defined for binary classification tasks with mutually exclusive labels. However, in this study, its logic was adapted to assess the performance of LLMs in a multi-label classification setting. Specifically, for each review, the relevance scores assigned by an LLM to individual issues are binarized: if a relevance score is greater than 0, it is converted to 1, indicating that the issue is present in the review; otherwise, it is set to 0.

To compute inter-model agreement, Cohen's Kappa is first calculated issue-wise between two models (e.g., LLM_x and LLM_y) by comparing their binary labels for each specific issue. This yields a separate Kappa score for each issue. Finally, an average Cohen's Kappa is computed across all issues to obtain an overall measure of agreement between the two models' multi-label classifications.

The standard formula for Cohen's Kappa is provided in Equation 8, while the adapted version used in this study (averaged across issues) is given in Equation 9.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{8}$$

$$\bar{\kappa}_{LLM_x, LLM_y} = \frac{1}{m} \sum_{j=1}^{m} \kappa_j \tag{9}$$

In these equations, p_0 is the obeserved agreement (the proportions of items where the two raters agree), p_e is the expected agreement by cance, computed from the inividual label distributions of the two raters, κ ranges from -1 (all $\kappa < 0 \Rightarrow$ agrement worse than chance) to 1 (perfect agreement) and represents the Cohen's Kappa score for one issue ($\equiv \kappa_j$ in equation 9), m is the total number of issues of LLM_x, and finally $\bar{\kappa}_{LLM_x,LLM_y}$ is the average Kappa measuring the overall agreement between LLM_x and LLM_y.

3.3.7 Krippendorff's Alpha

Krippendorff's Alpha is the third metric for inter-rater reliability used in this paper to asses the degree of agreement between the LLMs. It generalizes Cohen's Kappa metric and is particularly valued for its flexibility. Unlike many alternative measures, Krippendorff's Alpha supports any number of raters, missing data, and a variety of measurement levels, including nominal, ordinal, interval, and ratio [29]. This makes it especially suitable for evaluating complex classification tasks, such as the multi-label outputs of LLMs.

In this study, Krippendorff's Alpha was used to evaluate the consistency of issue classification across different LLMs. Unlike the adaptation needed for Cohen's Kappa, no binarization was required here. Since the LLMs produced relevance scores as continuous values (ranging from 0 to 1), the interval measurement level was adopted. This approach allows the metric to take into account the magnitude of disagreement, not just its presence, thereby offering a more nuanced evaluation. For example, two models assigning 0.85 and 0.9 relevance to the same issue are treated as more similar than models assigning 0.85 and 0.1, which would not be possible with binary or nominal encodings. As such, the use of the interval scale better reflects the nature of the LLM output and improves the interpretability of agreement.

For each pair of LLMs (e.g., LLM_x and LLM_y), Krippendorff's Alpha was computed issue-wise, and the average alpha score across all shared issues was used as the overall agreement score for that model pair. The metric is defined in equation 10, where D_o is the observed disagreement and D_e is the expected disagreement by chance.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{10}$$

An alpha of 1 indicates perfect agreement, 0 indicates chance-level agreement, and negative values indicate systematic disagreement. This setup allows Krippendorff's Alpha to serve as a statistically grounded, scale-sensitive measure of agreement between LLMs in a multi-label, interval-valued classification setting.

4 Data Analysis

4.1 Datasets Characteristics

This study focuses on identifying and analyzing user-reported issues in four Dutch government applications: KopieID, Reisapp, MijnOverheid, and DigiD. All four apps are publicly available on Google Play Store and contain user-submitted written reviews, accompanied by a star rating and the timestamp of each submission. To facilitate the analysis, user review data was collected using Google Play Scraper, resulting in one dataset per application.

These four applications serve distinct yet complementary functions within the Dutch government digital ecosystem. KopieID allows users to create a secure, anonymized copy of their identity documents, helping to prevent identity theft in online and offline processes [39]. Reisapp offers real-time travel advice, safety alerts, and essential information for Dutch citizens traveling abroad, assisting with consular services and travel preparations [40]. MijnOverheid acts as a centralized digital mailbox, providing users with official government correspondence, documents, and personal records in one accessible platform [38]. Lastly, DigiD functions as a national digital identification system, allowing citizens to securely access a variety of public and private sector services online, including tax filings, healthcare portals, and municipal services [37].

Each of the scraped datasets includes both the star ratings and the corresponding free-text reviews. Table 2 presents key statistics for the raw datasets, including the total number of reviews (i.e., reviews with and without written text), the number of scraped reviews (i.e., those that include text), the average star rating for each application, and the range of review dates (from the earliest to the most recent). It is important to note that these statistics reflect the data collected as of January 2025. Any subsequent execution of the scraping script may yield slightly different results due to new reviews being added or existing ones being removed from the platform.

	KopieID	**Reisapp**	**MijnOverheid**	**DigiD**
Google Play reviews	3.28k	8.15k	78.1k	306k
Google Play stars	1.9	3.8	4.4	4.3
Scrapped reviews	417	327	1056	4163
first review	Nov 2014	Jun 2012	Oct 2018	Mar 2017
last review	Jan 2025	Sep 2024	Jan 2025	Jan 2025

Table 2: Statistics of the raw datasets used, according to Google Play and Google Play Scraper. The information was extracted in January 2025.

Moreover, to accommodate the different characteristics and requirements of the involved models, two distinct preprocessing pipelines were developed: one tailored for preparing inputs for LLMs and another specifically designed for the LDA algorithm. LLMs, due to their scale and pretrained capabilities, can handle a degree of linguistic variability and noise. However, to maximize consistency, ensure clarity in prompts, and enable a fair comparative analysis between the LLMs and LDA, the preprocessing pipeline for LLMs included steps such as translation to English and removal of special characters. The detailed steps of this preprocessing workflow are outlined in one of the following sections.

In contrast, LDA is a probabilistic topic model that relies heavily on the statistical co-occurrence of terms within a corpus [4]. As such, its performance is highly sensitive to textual noise and irrelevant tokens. The LDA-specific preprocessing pipeline applied a more rigorous cleaning process besides the translation, including lemmatization, and stopword removal. The complete preprocessing pipeline will be presented in detail throughout the remainder of this chapter.

This dual-preprocessing approach ensures that both model types receive inputs that are best aligned with their structural assumptions and processing capabilities, thereby enhancing the validity and interpretability of the comparative evaluation. Table 3 presents an overview of the preprocessed datasets, including the number of reviews that were originally written in English and those that were translated into English during preprocessing.

	KopieID	**Reisapp**	**MijnOverheid**	**DigiD***
non-English reviews	248	169	723	2294
Dutch reviews	212	139	521	1582
English reviews	168	156	329	1843

** For DigiD, two reviews contained only numbers, no translation was needed for those.*

Table 3: Preprocessed Datasets statistics. The non-English reviews contain mostly Dutch reviews. All these reviews were translated to English using `googletrans 4.0.0rc1` Python package.

4.2 Preprocessing steps for LLMs

LLMs typically have built-in tokenization specifically designed to match the pre-trained embeddings of the LLM, which allows them to handle natural language directly [10], [33]. As a result, their preprocessing requirements are often simpler compared to traditional approaches.

1. Keep only the informative columns from the raw dataset, namely: `content` (the actual written review), `at` (the date on which the review was uploaded), and `appVersion`;
2. Clean the dataset by removing URLs, special characters (typically emojis), while preserving ”.”, ”,”, and ” ’” to maintain sentence structure and logic. Additionally, eliminate any extra white spaces;
3. Employ `googletrans` Python package to translate the non-english reviews to English.

4.3 Preprocessing steps for LDA

The preprocessing steps required to prepare the datasets for LDA are more extensive compared to LLMs. This is because LDA relies on the precise identification of word distributions within the text to extract meaningful topics [4].

1. Keep only the informative columns from the raw dataset, namely: `content` (the actual written review), `at` (the date on which the review was uploaded), and `appVersion`;
2. Clean the dataset by converting all text to lower case, removing the URLs, all special characters, any digits and extra white spaces;
3. Employ `googletrans` Python package to translate the non-english reviews to English;
4. An additional cleaning phase was implemented to address special characters, such as “’” and capital letters, introduced during the translation process;

5. Handle negation phrases, such as “not working”, by converting them into “not_working.” This ensures the user’s intent is preserved when describing a topic and prevents the negation (“not”) from being removed in subsequent preprocessing steps;
6. Tokenize the reviews;
7. Remove the stop words using the `nltk` package;
8. Use lemmatization to reduce words to their base form and standardize the vocabulary for consistency.

Observation 2:

The number of preprocessed reviews used for LDA may differ from those used for the LLM. This is because LDA preparation involves more extensive text cleaning, particularly after translation and tokenization. During this process, reviews that consist solely of stop words are entirely removed, resulting in empty entries that are subsequently discarded.

Observation 3:

In the initial phase of this study, individual star ratings were not considered particularly relevant to the analysis. However, it later became evident that these ratings provide valuable insights and should be included in the evaluation. To integrate the preprocessed reviews with their corresponding original star ratings, a merge operation was performed based on the review timestamp, including both the date and specific hour. Additional checks were carried out to ensure that each preprocessed review was matched uniquely to a single original review, maintaining the integrity of the data alignment.

5 Experimental Results

5.1 Issue Extraction

The first part of this analysis focuses on extracting topics or issues from application reviews using both LDA and five different LLMs: GPT-4o-mini, Cluade-3.5-Sonnet, Gemini-1.5-Pro, Gemini-2.0-Flash, and Mistral-Large-2411.

This section begins with an overview of the overall scores, an analysis of topic similarities compared to the traditional model, and an exploration of how issues identified by different LLMs are interconnected.

All LLMs were prompted using the same instruction, which asked them to extract the types of issues present in a dataset of preprocessed reviews (see Section 4), along with a list of the most relevant words associated with each identified issue. These relevant words play a crucial role in the subsequent evaluation, as they are used to compute the semantic similarity between issue sets generated by different LLMs, and also between each LLM and the LDA model. This is achieved by constructing TF-IDF embeddings from the extracted words and calculating their cosine similarity.

To enable a fair and meaningful comparison, the extracted relevant words needed to be uniformly processed, as LLMs may output multi-word expressions (e.g., "language barrier") that require further refinement. Therefore, a dedicated cleaning pipeline was applied to the extracted word lists, which included the following steps: splitting multi-word expressions into individual words, handling negations in a manner consistent with LDA preprocessing, tokenization, removal of stop words, and elimination of duplicate words. This standardization process ensured that the resulting word embeddings were comparable across models, allowing for a robust evaluation of the coherence and similarity of the extracted issues.

Table 4 presents a summary of statistics regarding the number of issues (Num. issues) extracted by each LLM, as well as the average number of cleaned relevant words per issue (Avg. words). Overall, Gemini-2.0-Flash produces the highest number of relevant words per issue, followed by Mistral-Large-2411, suggesting that these models tend to generate more descriptive or elaborated keyword sets.

	Metric	KopieID	Reisapp	MijnOverheid	DigiD
gpt-4o-mini	Num. issues	10	8	10	8
	Avg. words	9.40	8.00	10.20	10.88
claude-3.5-sonnet-20241022	Num. issues	8	7	8	6
	Avg. words	10.13	8.86	8.88	11.50
gemini-1.5-pro	Num. issues	7	6	9	8
	Avg. words	11.00	8.17	7.22	13.13
gemini-2.0-flash	Num. issues	9	8	5	13
	Avg. words	16.89	10.13	20.20	10.15
mistral-large-2411	Num. issues	10	10	13	7
	Avg. words	9.70	11.20	13.23	18.86

Table 4: Number of issues and the average word count per issue for each LLM and application

When it comes to the number of extracted issues, there is no consistent agreement across models for any of the applications. For instance, in the case of MijnOverheid, the number of identified issues ranges from 5 (Gemini-2.0-Flash) to 13 (Mistral-Large-2411). This variation highlights the diverse levels of granularity employed by different LLMs. Some models tend to provide a more specific and fine-grained categorization, while others capture broader thematic concerns, potentially prioritizing generalization over detail.

5.1.1 Coherence Scores

The next stage in the analysis involves evaluating the quality of the extracted issues using coherence scores, which serve as a measure of the semantic consistency among the relevant words grouped under each issue. The results of this evaluation are presented in Table 5. The highest coherence scores were observed for Gemini-2.0-Flash on KopieID and MijnOverheid, Claude-3.5-Sonnet on Reisapp, and Mistral-Large-2411 on DigiD. Notably, across all four applications, LDA was consistently outperformed by the majority of the LLMs, highlighting the superior topic modeling capabilities of modern language technologies in this context.

A comparison of Table 4 and Table 5 reveals a notable pattern: LLMs that produce a larger number of cleaned relevant words per issue tend to achieve higher coherence scores. This is particularly evident in the cases of Gemini-2.0-Flash and Mistral-Large-2411, which both exhibit high word counts and top coherence performance. As the number of semantically related words increases, the model has more opportunity to demonstrate internal consistency, resulting in a higher overall score. Thus, these models not only perform well but can also be considered more trustworthy in the task of issue extraction.

Model	**Overall Coherence Score**			
	KopieID	**Reisapp**	**MijnOverheid**	**DigiD**
LDA_5_topics	0.3164	0.2920	0.4409	0.4658
LDA_7_topics	0.3330	0.3489	0.4019	0.4575
LDA_10_topics	0.3189	0.3435	0.4137	0.4541
gpt-4o-mini	0.4318	0.5106	0.4295	0.3964
claude-3-5-sonnet-20241022	0.4650	0.5658	0.3906	0.4827
gemini-1.5-pro	0.3401	0.5063	0.4395	0.4284
gemini-2.0-flash	0.4710	0.5343	0.5303	0.5549
mistral-large-2411	0.4673	0.5429	0.5224	0.5564

Table 5: Comparison of Overall Coherence Scores Across Applications and Models. For each application, the two highest coherence scores among the LLMs are highlighted: the highest, in green and the second-highest, in yellow.

> Across all four applications, the majority of LLMs consistently achieved higher coherence scores than LDA, underscoring the enhanced ability of modern language technologies to extract well-structured and meaningful topics from user feedback.

5.1.2 LDA vs LLMs

In the previous subsection, the overall differences between LDA and the LLMs were presented, concluding that LDA generally achieves lower coherence scores, especially when considering the varying number of topics. However, beyond this performance gap, it is also important to examine how the outputs of these two types of models relate to one another in terms of the issues they identify. Specifically, this raises the question: *to what extent are the issues discovered by LLMs also captured by LDA?* To explore this, several similarity heatmaps were generated, which will be presented and analyzed throughout this section.

To streamline the analysis and ensure clarity, only the LDA model with 10 topics was selected for these comparisons. This decision was based on the observation that coherence scores were relatively consistent across the LDA models with 5, 7, and 10 topics, and using the highest-topic model allows for a broader issue coverage, improving the interpretability of the comparisons.

In the next step, for each application, a heatmap was analyzed comparing the LDA model with 10 topics and the LLM that achieved the highest coherence score (according to Table 5) for that specific application. This decision was made to reduce and concentrate the analysis, avoiding an overload of visualizations while ensuring that only the most relevant and high-performing models are compared.

KopieID

For KopieID, the heatmap presented in Figure 2 illustrates the topic similarity between LDA with 10 topics and the Gemini-2.0-Flash, which was selected due to its highest coherence score. The heatmap reveals that, although Gemini-2.0-Flash and LDA do not share perfect alignment, several points of notable similarity emerge. Specifically, LDA Topic 1 shows the strongest correspondence with Gemini's "Scanning Issues", displaying a similarity score of 0.57, indicating a clear overlap in how both models capture this prominent issue in the reviews. Two additional topics from LDA, specifically Topics 7 and 8, also exhibit clear correspondences with the issues extracted by Gemini-2.0-Flash, aligning with the "Image Quality Issues" and "Incorrect Masking/Redaction" categories, respectively.

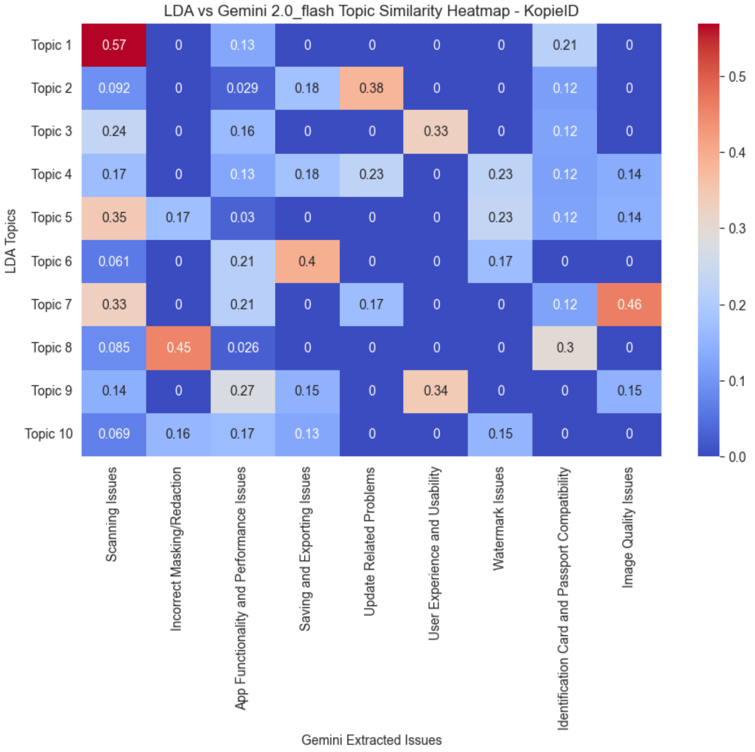


Figure 2: Topic similarity heatmap for KopieID

Another important observation is that some issues identified by the LLM exhibit distributed correspondences across multiple LDA topics, rather than aligning strongly with a single topic. In these cases, the similarity scores are spread over several LDA topics, typically with lower individual values. For instance, the issue "App Functionality and Performance Issues", which represents a broader and more general category, shows moderate similarity scores across multiple LDA topics. Similarly, more specific issues such as "Watermark Issues" and "Identification Card and Passport Compatibility" also correspond to several LDA topics, though with lower individual scores.

This pattern suggests that the LLM's issue extraction tends to capture more nuanced and multidimensional themes, integrating aspects that LDA, with its rigid topic structure, distributes across separate topics. From an analytical perspective, this can be seen as an advantage of LLMs: rather than fragmenting related concerns across isolated topics, LLMs consolidate semantically connected aspects under broader or more coherent issue labels. This makes the results more interpretable, especially for practitioners looking to address complex, multifaceted problems reported by users.

Overall, this analysis reinforces the strength of Gemini-2.0-Flash in capturing user-relevant issues with greater coherence and specificity, while still maintaining meaningful overlaps with LDA baseline.

Reisapp

Figure 3 illustrates the similarity between the LDA model with 10 topics and Claude-3.5-Sonnet model, which achieved the highest coherence score among all tested LLMs for this application.

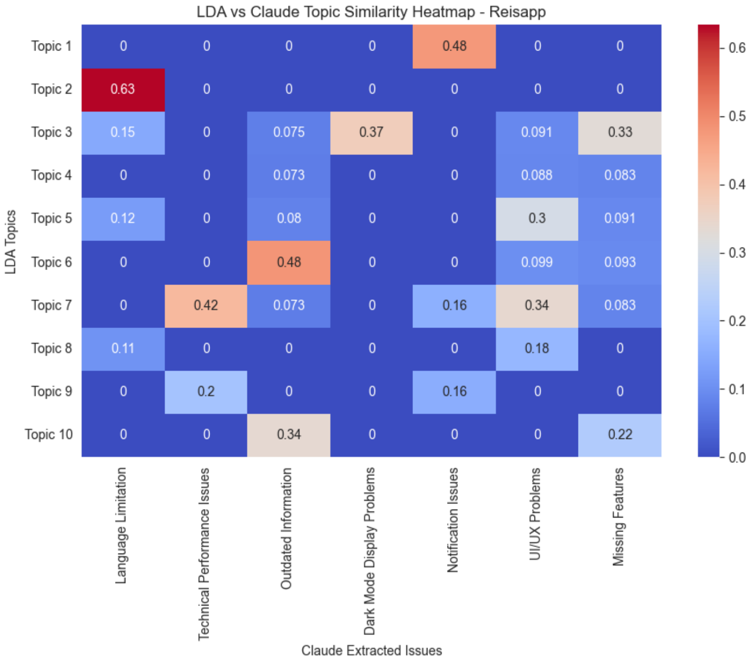


Figure 3: Topic similarity heatmap for Reisapp

The most prominent alignment is observed between LDA Topic 2 and Claude's "Language Limitation", with a similarity score of 0.63, the highest across the entire matrix. This indicates that both models consistently identified the lack of multilingual support as a significant concern among Reisapp users. Furthermore, LDA Topic 6 shows a meaningful alignment with "Outdated Information", scoring 0.48, reflecting user frustrations with obsolete or inaccurate travel updates. Another notable overlap appears with the application's notification system, where "Notification Issues" from Claude corresponds to LDA Topic 1, underlining that both models captured concerns related to excessive or ineffective notifications.

An interesting detail emerges with Claude's "Dark Mode Display Problems", which exhibits alignment solely with LDA Topic 3, without spreading across multiple LDA topics as observed with other issues. This suggests that the visual and accessibility problems related to dark mode are recognized in a more isolated and coherent way by both models. In other words, the concerns around dark mode functionality appear to be well-contained and distinctly recognized within user feedback, rather than being part of a broader, entangled issue cluster. This clear one-to-one alignment might indicate that users express dark mode problems in consistent and specific terms, making it easier for both LDA and Claude to map these complaints under a single thematic category.

Furthermore, in the case of Reisapp, certain issues identified by Claude, such as "UI/UX Problems" and "Missing Features", do not correspond strongly to a single LDA topic but instead display moderate similarities spread across multiple LDA topics. This pattern indicates that Claude-3.5-Sonnet, similar to Gemini-2.0-Flash, is capable of detecting broader, cross-cutting themes within the user feedback, grouping together aspects that LDA's topic modeling approach tends to separate. Such an outcome demonstrates Claude's ability to provide a more integrated perspective on user concerns, which can be particularly valuable for decision-makers.

MijnOverheid

For the third studied application, the heatmap presented in Figure 4 shows the relationship between the LDA model with 10 topics and the Gemini-2.0-Flash model, which once again delivered the highest coherence score.

The strongest correspondence emerges between LDA Topic 9 and Gemini's "DigiD Login Issues", with a similarity score of 0.48. This highlights a significant concern among users, given that MijnOverheid heavily depends on DigiD for secure identification and access to governmental services. The critical role of this functionality makes its prominence in user feedback particularly notable.

Further, relevant overlaps include LDA Topic 5 or 6, which corresponds to Gemini's "Content and Coverage Limitations". These scores reflect recurring user complaints about insufficient information availability or restricted access to essential services within the platform.

Similarly, LDA Topic 7 shows moderate similarity (0.40) with "User Interface and Experience Issues", indicating users' frustrations with the app's usability and design aspects.

What stands out in this case is that Gemini's "General Dissatisfaction" category does not map strongly to any single LDA topic but instead displays weaker, scattered similarities across multiple topics. This dispersion suggests that general dissatisfaction likely stems from a combination of smaller, less isolated issues, which LDA struggles to unify under one topic. In contrast, the LLM successfully groups these fragments into a cohesive category, offering a hint on the user sentiments.

Thus, this heatmap reinforces Gemini-2.0-Flash's ability to capture both specific technical problems and more diffuse, user-experience-related frustrations.

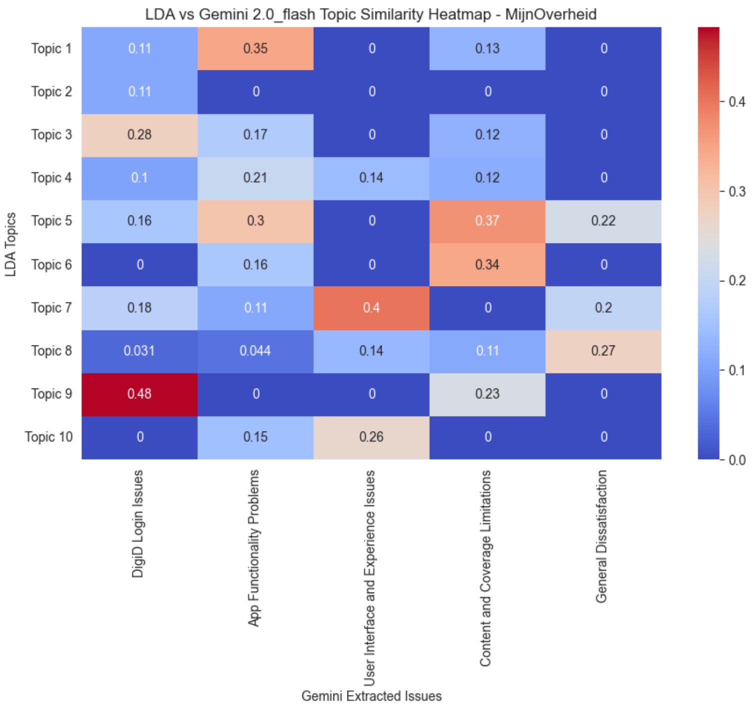


Figure 4: Topic similarity heatmap for MijnOverheid

DigiD

For DigiD, the heatmap in Figure 5 shows the alignment between the LDA model with 10 topics and the issues extracted by Mistral-Large-2411. Compared to the previous applications, DigiD's heatmap reveals a more distributed pattern of similarities, with fewer extremely dominant matches, but several moderate correspondences that still provide meaningful insights.

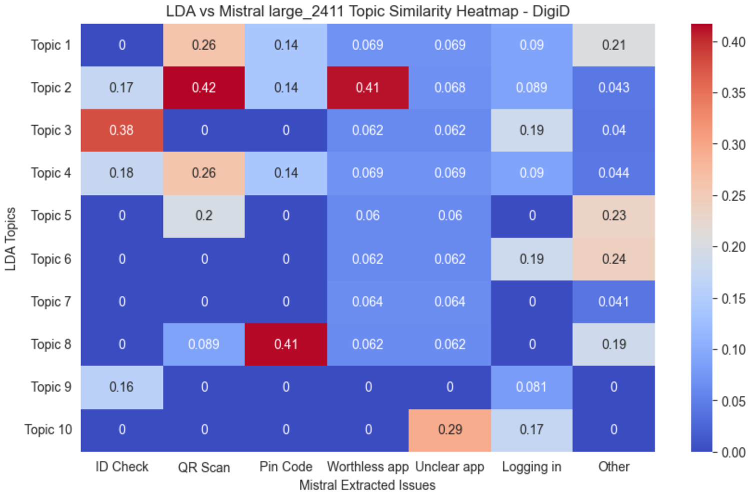


Figure 5: Topic similarity heatmap for DigiD

The most notable alignment is between LDA Topic 2 and Mistral's "QR Scan" issue, with a similarity score of 0.42, the highest in this heatmap. This suggests that both models identified technical problems related to the QR scanning functionality as a significant pain point for DigiD users. Closely following, LDA Topic 2 also aligns strongly with "Worthless app", scoring 0.41, indicating a cluster of user dissatisfaction that may not target a specific technical failure but rather a broader sense of frustration or disappointment with the app's usefulness. Another important observation is the similarity between LDA Topic 8 and "Pin Code" issues. This highlights user concerns around security or usability aspects related to PIN code handling — a critical feature in an authentication app like DigiD.

Interestingly, the "Other" category from Mistral shows a more dispersed similarity across several LDA topics, albeit with lower individual scores. This dispersion suggests that miscellaneous issues, which do not neatly fit into predefined categories, are scattered throughout the LDA topics — highlighting again the LLM's strength in capturing subtle, cross-cutting themes that LDA tends to fragment. However, this observation also points to a potential limitation of Mistral-Large-2411: by clustering too many heterogeneous complaints under a broad "Other" label, the model may risk losing clarity and specificity, making it harder for practitioners to extract precise insights from this particular category. While this flexibility enables the model to recognize less frequent or ambiguous user concerns, it also introduces some interpretability challenges when trying to differentiate between distinct, lower-frequency issues.

Overall, while the similarities in this heatmap are somewhat more spread out compared to previous applications, Mistral-Large-2411 demonstrates its ability to capture both specific technical issues (like QR scanning and PIN code problems) and more generalized user frustrations. This mixed pattern underlines the model's versatility in handling both focused concerns and broader, less well-defined dissatisfaction.

> The analysis demonstrates that LLMs consistently outperform LDA in extracting coherent and interpretable issues from user reviews, offering broader thematic coverage and effectively capturing cross-cutting concerns that LDA tends to fragment. Their flexibility enables LLMs to consolidate related problems under clearer categories, providing richer insights for decision-makers. However, this same flexibility can occasionally result in overly broad groupings, which may reduce interpretability for less frequent or ambiguous issues. In contrast, LDA, while more rigid and limited in coherence, provides clearer separation between topics, which can help pinpoint specific issues without overlap — although it lacks the ability to explicitly name or label the topics, leaving interpretation to the analyst.

5.1.3 Issue Clusters

In the previous section, various LLMs were compared with the traditional LDA to examine the alignment of extracted issues across approaches. In this section, the focus shifts toward exploring the commonalities among the LLMs themselves: identifying shared issues, understanding their nature, and investigating whether they can be grouped into broader thematic categories. To achieve this, cosine similarity was employed to measure the closeness between issues extracted by different LLMs. Specifically, for each issue I_a from LLM_x, connections were established to the most similar issues from every other LLM_y, where $x \neq y$. The resulting structure forms a graph (see Figures 6-9), where each node represents an issue, labeled as $LLM_x : I_a$. To identify meaningful groupings within this network, the Louvain algorithm was applied, allowing for the detection of clusters of related issues within each application, and thereby revealing overarching themes shared across different LLM outputs. These clusters are visually distinguished by distinct colors in the graphs, enhancing interpretability and making the relationships between grouped issues more evident.

To provide a clearer understanding of the shared issues and their thematic groupings, the analysis will examine each application individually, presenting its corresponding cluster graph. For this analysis, only GPT-4o-mini, Gemini-1.5-Pro, Gemini-2.0-Flash, and Mistral-Large-2411 were included, as Claude-3.5-Sonnet is also omitted from the subsequent two main tasks.

KopieID

The cluster graph presented in Figure 6 for KopieID reveals thirtheen well-formed groups of related issues, demonstrating that across different LLMs, there is considerable alignment in the types of concerns identified by users.

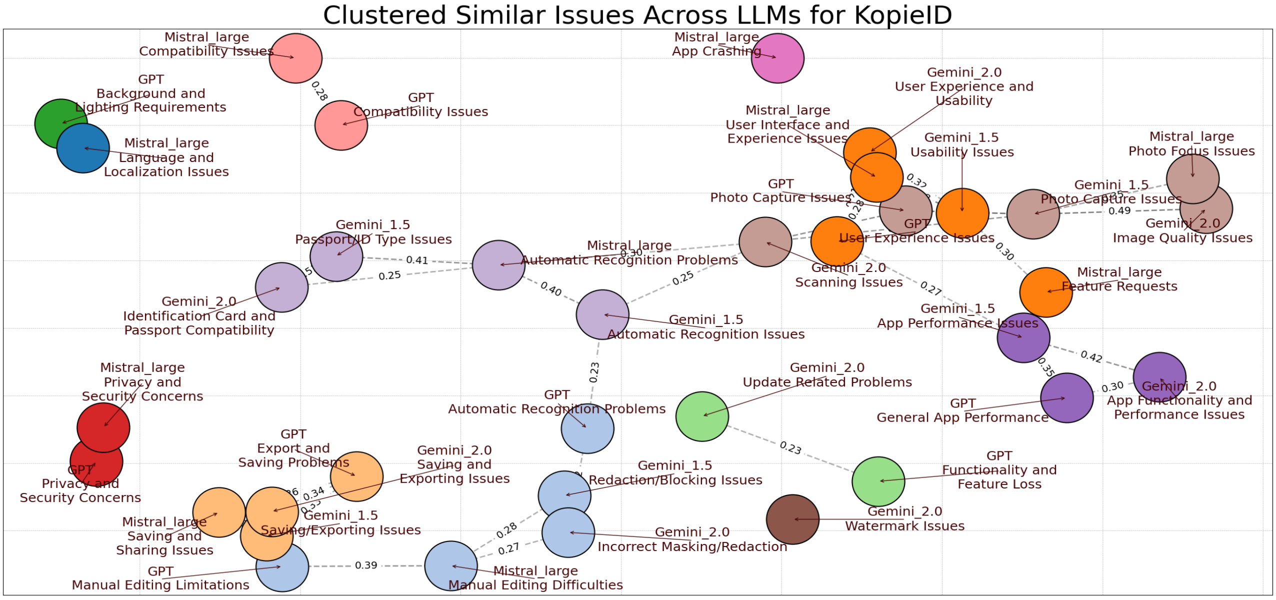


Figure 6: KopieID issue clusters according to 4 LLMs

The primary clusters detected — all consisting of more than two interconnected nodes — are associated with the following categories of issues:

Scanning and Photo Quality Issues ● Issues such as "Photo Capture Issues", "Image Quality Issues", "Photo Focus Issues", and "Scanning Issues" are grouped together, involving all four LLMs. The cohesion of this cluster highlights persistent user challenges related to document capture quality, which is crucial for an app whose main purpose is to anonymize documentation by taking photos.

Passport and ID Recognition Issues ● This cluster brings together concerns such as "Identification Card and Passport Compatibility", "Passport/ID Type Issues", and both "Automatic Recognition Issues" and "Automatic Recognition Problems". The presence of multiple models in this cluster signals a strong cross-model agreement that document recognition remains a critical challenge in KopieID. Users appear to frequently encounter problems with the automatic recognition of identity documents, whether due to incompatibility with certain document types or technical failures in scanning and verification. Moreover, this cluster is also strongly connected to the Scanning and Photo Quality one.

Saving and Exporting Issues ● This cluster groups together issues like "Saving/Exporting Issues", "Saving and Sharing Issues", and "Export and Saving Problems". The coherence of this cluster suggests that users of KopieID consistently struggle with functionalities related to saving, exporting, and sharing their documents. Whether due to unclear export procedures, file format limitations, or technical glitches during the saving process, these issues appear to be closely related and frequently reported.

General Performance Issues ● Grouping terms such as "App Functionality and Performance Issues", "General App Performance", and "App Performance Issues", this cluster highlights operational reliability problems. Its spread across Gemini models and GPT suggests a shared perception of KopieID's technical performance limitations.

User Experience and Usability Issues ● This cluster contains issues like "User Experience Issues", "Usability Issues", and "User Interface and Experience Issues", cutting across models. It points to broader design and accessibility concerns, emphasizing that multiple LLMs recognized user frustration not just with technical problems but also how users interact with the app.

Manual Editing and Redaction Issues ● This cluster centers around challenges related to manual document editing and redaction functionalities, including issues like "Manual Editing Difficulties", "Manual Editing Limitations", "Redaction/Blocking Issues", and "Incorrect Masking/Redaction".

The clear focus of this cluster suggests that users expect greater precision and reliability from these editing features, which are especially important in an application handling personal identification documents. Problems in this area not only impact usability but also raise potential privacy and compliance concerns, if sensitive data cannot be properly obscured or removed.

Across the smaller clusters, typically comprising two nodes, the following groupings emerge:

Privacy and Security Issues ● Both GPT and Mistral-Large-2411 identified issues related to privacy and data security, which are tightly grouped in the red cluster. This strong alignment across models underscores the importance of proper handling of sensitive personal information in KopieID, a critical concern for users of an identification app. Notably, as previously discussed, these concerns can be connected to the Manual Editing and Redaction Issues, where difficulties in obscuring or removing sensitive data may further exacerbate privacy risks.

Compatibility Issues ● This cluster brings together concerns related to device compatibility and technical environment requirements. The appearance of this cluster highlights user frustrations stemming from difficulties in running KopieID smoothly across different devices, operating systems, or configurations. Such issues are critical for an identification application expected to work seamlessly across a wide range of user environments. The coherence of this cluster, despite its smaller size, underscores the need for robust cross-device support and careful attention to technical dependencies that may hinder accessibility.

Reisapp

For Reisapp, the clusters presented in Figure 7 appear to be much more clearly defined and exhibit considerably less interconnection compared to the previous application. In total, six major clusters have been identified, each containing at least three nodes. Interestingly, there are also five isolated clusters consisting of only a single node, indicating that certain issues were uniquely identified by individual LLMs. This suggests either highly specific concerns that were not widely recognized across models, or lower similarity scores that prevented integration into larger clusters. For the purposes of this analysis, the focus will remain on the main clusters, as they represent the most prominent and recurring themes identified across the models.

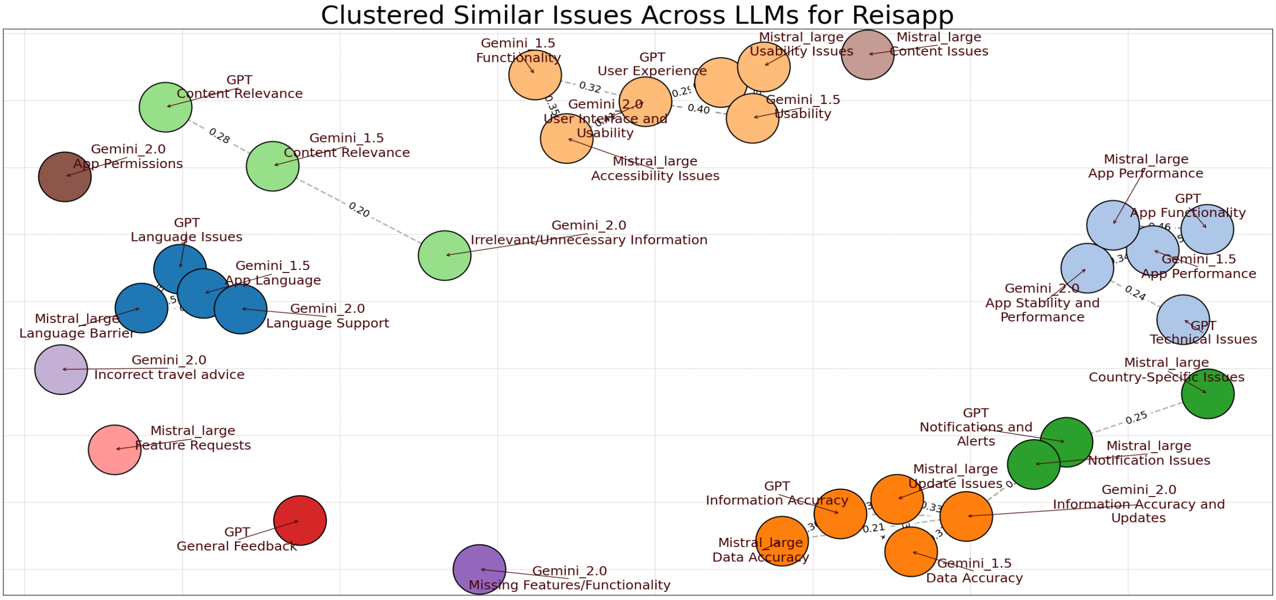


Figure 7: Reisapp issue clusters according to 4 LLMs

User Experience and Usability Issues ● This cluster is one of the largest and most cohesive in the graph, bringing together issues like “User Experience”, “Usability”, “Functionality”, “User

Interface and Usability", and "Accessibility Issues" across all LLMs. The high connectivity within this group indicates that users consistently report challenges related to the design and navigability of the app. This spans across general ease of use, clarity of interface elements, and accessibility concerns. The consensus among LLMs highlights this as a critical area for improvement, with strong potential to enhance user satisfaction if addressed.

App Stability and Performance Issues ● This cluster includes "App Performance", "App Stability and Performance", "App Functionality", and "Technical Issues". These issues point toward persistent technical shortcomings within Reisapp, from performance lag to potential app crashes or glitches. The models converge strongly here, reflecting widespread user concerns about the app's reliability.

Language Issues ● Comprising issues like "Language Issues", "App Language", "Language Barrier", and "Language Support", this cluster highlights user difficulties with multilingual support and accessibility for non-Dutch speakers. It signals a clear demand for improved language inclusivity, ensuring that Reisapp can serve a broader audience effectively. The presence of this cluster across multiple models suggests that language limitations are a recurring and recognized obstacle.

Information Accuracy and Updates Issues ● This group contains issues such as "Information Accuracy", "Update Issues", and "Data Accuracy". It reflects outdated or incorrect travel information and possible delays in updates. The cluster underscores the importance of providing accurate, real-time data in travel-related apps.

Content Relevance Issues ● This smaller but still notable cluster gathers "Content Relevance" and "Irrelevant/Unnecessary Information" issues. Users appear to find that some of the app's content does not align with their needs or expectations. This suggests an opportunity to streamline and personalize content delivery, ensuring that users receive relevant, concise, and actionable information.

Country-Specific and Notification Issues ● Finally, this cluster includes "Notification Issues" and "Country-Specific Issues", which points to a combination of concerns: how notifications are handled, and how well the app adapts to the specifics of different regions or countries. These issues likely reflect both technical and content customization challenges, emphasizing the importance of localization and context-aware notifications for improving user experience.

MijnOverheid

For MijnOverheid, the issue clustering shown in Figure 8 displays a graph that, while less structured than that of Reisapp, still offers valuable insights into the distribution and nature of user concerns. Specifically, five major clusters have been identified, each containing at least three interconnected nodes, indicating substantial alignment among the models on these themes. In addition to these primary groups, there are also two smaller yet noteworthy clusters.

Primary clusters:

Login and Authentication Issues ● This is one of the most dominant clusters in the graph, encompassing concerns like "Login Issues", "DigiD Login Issues", "DigiD App Issues", and "Sync and Update Issues". Given that MijnOverheid relies heavily on secure digital identification (using DigiD), it is unsurprising that access and login difficulties are top-of-mind for users. The cluster shows strong agreement across models, reinforcing that secure and stable login experiences are essential for user trust.

App Performance and Stability Issues ● This cluster covers issues such as "App Performance", "App Crashes/Freezes", "General Dissatisfaction", "Slow Performance", "Document Handling", and "Message Content". It reflects frustrations with the app's reliability, responsiveness, and how well it stays up to date. The presence of general dissatisfaction within this cluster underscores how technical shortcomings strongly influence overall user sentiment. Notably, this cluster does not exist in isolation: it is closely connected to the previously discussed Login and Authentication Issues, as technical instability can directly impact access reliability. Furthermore, its links to the following User Interface and Experience cluster highlight the interdependence between technical performance and user perception of app usability.

User Interface and Experience Issues ● In this case, issues like "User Interface Problems", "User Interface and Experience Issues", "User Experience (UX) Issues", "Display and Layout Issues", and "Usability Issues" can be observed. The concentration of these concerns into a cohesive cluster shows that users desire a more intuitive and aesthetically pleasing design. Problems range from confusing layouts to poor usability, underlining the importance of user-centered design principles.

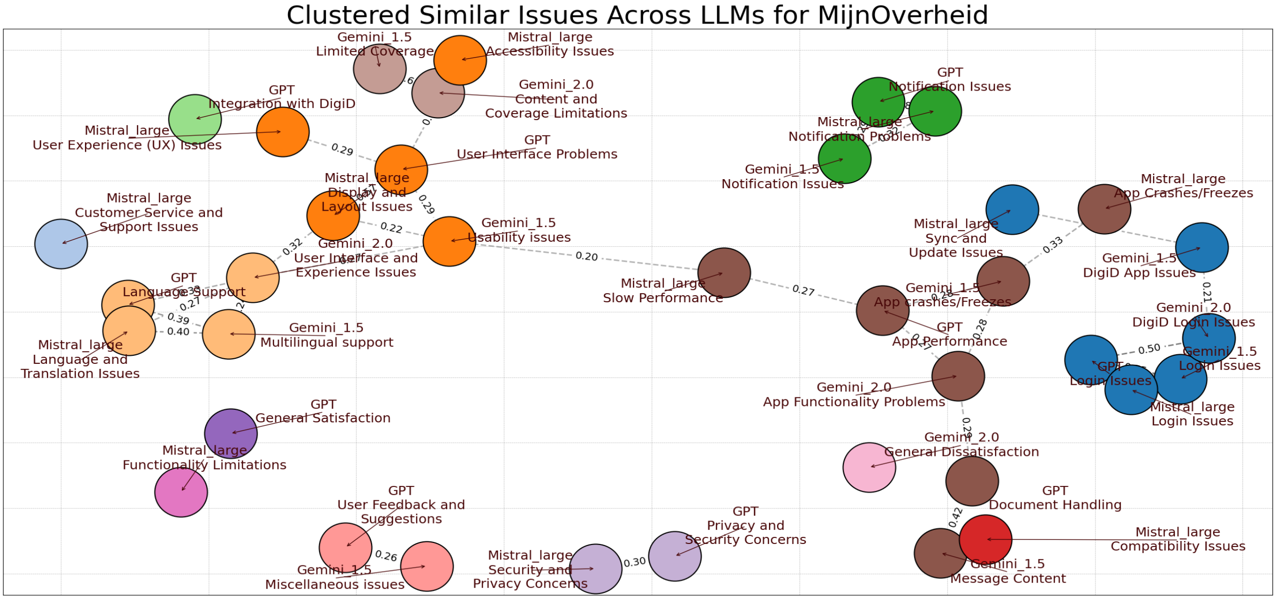


Figure 8: MijnOverheid issue clusters according to 4 LLMs

Language Support and Accessibility Issues ● This cluster gathers issues such as "Language Support", "Multilingual Support", and "Language and Translation Issues". As MijnOverheid serves a wide population, including non-native speakers, the presence of this cluster signals the necessity of robust language options and clear translations.

Notification Issues ● This cluster focuses on concerns related to the app's notification system, including issues like "Notification Problems" and "Notification Issues". This cluster plays a significant role given the app's function in delivering important government updates and reminders to users. The concentration of these issues signals user frustration with notifications that might be either delayed, unreliable, or inconsistent in delivery.

Smaller clusters:

Privacy and Security Issues ● The graph shows a small but compact cluster, linking "Privacy and Security Concerns" and "Security and Privacy Concerns". The presence of these concerns, even in a more compact cluster, is particularly significant given the sensitive personal data managed by MijnOverheid. Users are clearly attentive to how their data is handled, and any perceived weaknesses in security protocols or communication can lead to a rapid erosion of trust.

Coverage Limitations Issues ● This cluster includes "Content and Coverage Limitations" and "Limited Coverage", forming a tidy, isolated group. There are a couple of weaker links to the User Interface cluster, hinting that content limitations are not purely data problems but can also influence users' perceived usability. As before, this suggests the importance of content completeness and accessibility in building trust and utility for diverse user segments.

DigiD

The clustering analysis for DigiD reveals a diverse and insightful landscape of user concerns, with six prominent clusters emerging from the graph presented in Figure 9.

Login and Verification Issues ● In this cluster issues like "Login Problems", "SMS Verification Problems", "Pin Code", and "Account Issues" are presented. Within this group, this type of problems emerge as interconnected concerns. This indicates that users specifically perceive the verification and authentication flow as its own significant pain point. This cluster highlights the challenges users face within DigiD's already complex authentication process, emphasizing the need to improve the clarity and reliability of multi-factor authentication and recovery procedures.

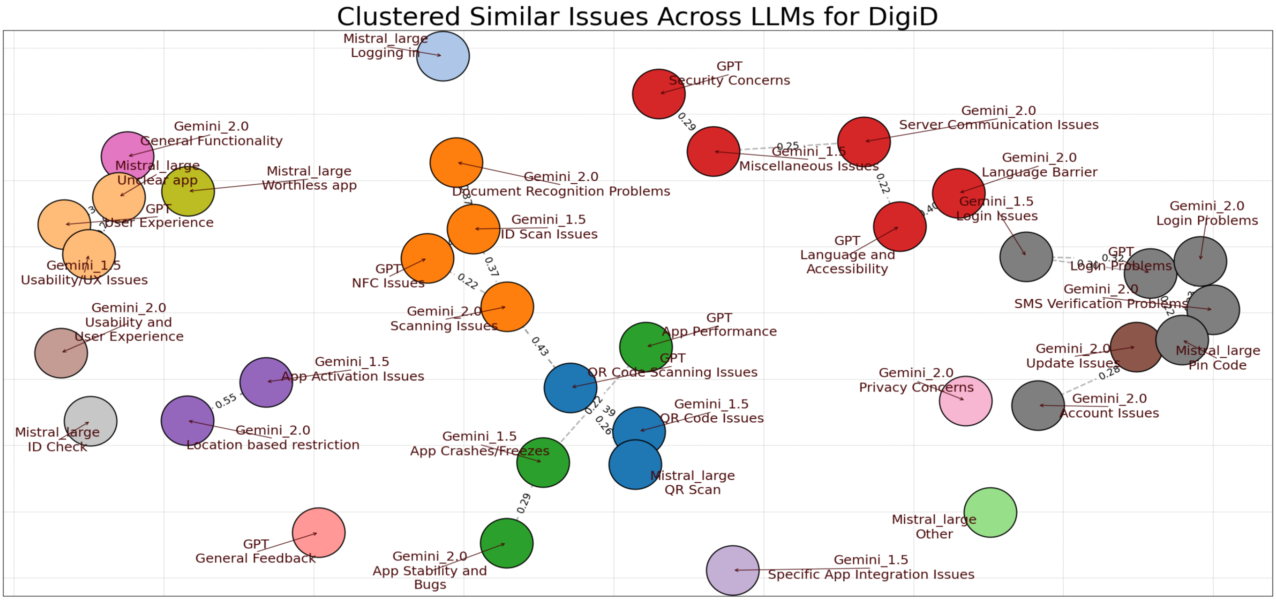


Figure 9: DigiD issue clusters according to 4 LLMs

App Performance and Stability Issues ● The green cluster centers on technical concerns related to the overall performance and stability of the DigiD application. It includes issues such as "App Performance", "App Crashes/Freezes", and "App Stability and Bugs", all pointing to user frustrations with the app's reliability during regular use. This cluster reflects general operational failures that undermine trust in the application as a dependable service.

QR Code Issues ● This cluster revolves around DigiD's scanning functionalities, including concerns such as "QR Code Scanning Issues", "QR Code Issues", and "QR Scan". This cluster reflects users' frustrations with the app's core identification mechanisms, where failures in scanning or recognizing identity QRs directly block the successful completion of critical tasks.

Scanning and Verification Issues ● This cluster, while closely connected to the QR Code Issues cluster, extends beyond it by bringing together a broader range of concerns related to scanning and verification functionalities. It encompasses issues such as "Document Recognition Problems", "ID Scan Issues", "Scanning Issues", and "NFC Problems", indicating that user frustrations are not limited to QR-related processes but span the entire document verification flow. The grouping of these issues highlights recurring technical weaknesses in the app's core identification functionalities. Failures in accurately reading identity documents or establishing Near Field Communication (NFC) connections often prevent users from completing essential verification.

User Experience and General Functionality Issues ● This cluster captures general user experience concerns, blending usability frustrations, app clarity problems, and broader user experience complaints. Unlike the technically focused clusters, this one reflects qualitative dissatisfaction with the overall usefulness and intuitiveness of DigiD. Its presence signals that beyond technical fixes, attention should also be directed toward improving user guidance, interface design, and the perceived value of the app.

Mixed Critical Issues (Security, Language, Communication) ● The most critical cluster is the red one, which presents a diverse mix of issues such as "Security Concerns", "Language Barriers", “Server Communication Problems”, and some miscellaneous complaints. Although these issues might seem unrelated at first glance, they share a common underlying factor: they all relate to fundamental barriers in accessing or trusting the application. Server errors, language barriers, and data security concerns all undermine the app’s intended functions. This convergence suggests that the stability of DigiD’s infrastructure, its multilingual accessibility, and its transparent communication regarding security are perceived as deeply intertwined by users.

> The clustering analysis shows that LLMs generally converge on the main user issues, though some models surface unique concerns. Across government applications, the most common issues relate to *technical problems* such as authentication failures, scanning errors, and notification issues. In addition, *security concerns* and the *lack of multilingual support* consistently appear, alongside broader themes of *poor user experience* and *app performance limitations*. These findings highlight the need to address both technical reliability and user-centric improvements to enhance overall satisfaction.

5.2 Review Classification

In this section, a multi-label classification is performed across all four applications, based on the issues extracted in the previous section. The classification is carried out using four LLMs: GPT-4o-mini, Gemini-1.5-Pro, Gemini-2.0-Flash, and Mistral-Large-2411.

Model performance was first assessed using Shannon entropy, applied across all LLMs and LDA for each application, to capture the diversity and uncertainty in topic assignments. These entropy plots, together with the coherence scores, form the foundation for determining the most effective model to be used in subsequent analyses, including forecasting as well as temporal and impact analyses.

As a second method for evaluating model performance, three agreement metrics were used to compare the consistency of LLM outputs. These metrics offer insight into how similarly the models classify reviews, highlighting their relative reliability.

5.2.1 Shannon Entropy

This section presents a consolidated figure for each NLP model, showing bar plots of Shannon entropy across all applications. This approach allows for a clearer comparison of how each model behaves across datasets of varying complexity and sizes, providing deeper insights into the models’ confidence and prediction diversity.

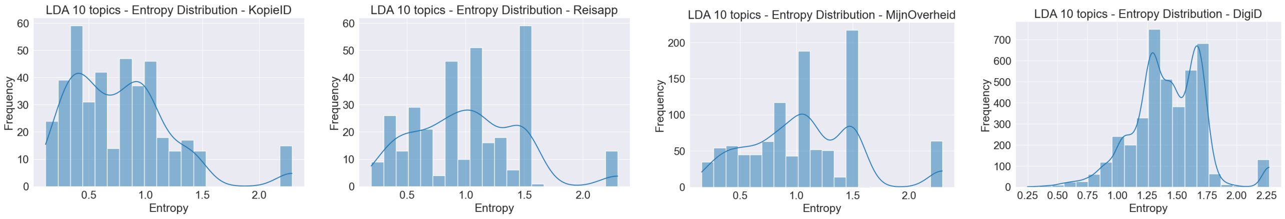
As a reminder, lower Shannon entropy indicates greater model confidence and clearer topic separation. Thus, models with lower entropy across applications are considered to perform better in this context.

LDA

The Shannon entropy analysis begins with an examination of the results produced by the traditional LDA model, as illustrated in Figure 10.

For KopieID, the entropy distribution displays a moderate range, with the majority of values hovering around 0.5–1.5. This suggests that while LDA is reasonably certain about the predominant topics in many reviews, there is nonetheless a meaningful spread of uncertainty. However, there are still a few reviews where the classification exhibits significant uncertainty, with entropy values exceeding 2.

Reisapp’s entropy values indicate less certain classifications compared to KopieID, with a peak around 1.5. Although the distribution still peaks around a mid-entropy value, the long tail exceeding again 2.0 highlights that certain reviews are notably difficult to classify under distinct topics, leading to higher uncertainty in those cases.



(a) KopieID (b) Reisapp (c) MijnOverheid (d) DigiD

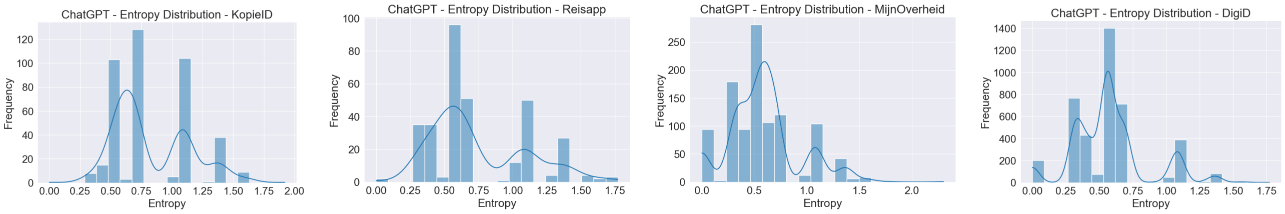
Figure 10: Shannon entropy across applications for LDA

MijnOverheid exhibits a similarly wide spread of entropy values. The elevated proportion of reviews in these middle-to-higher entropy ranges implies that LDA has some difficulty cleanly separating the issues within this dataset. This could reflect broader or more nuanced topics, where multiple issues overlap within the same review.

As the largest dataset, DigiD shows the most pronounced distribution at higher entropy values, with a cluster around 1.25–2.0 and some instances extending even beyond that. This pattern underscores a substantial degree of uncertainty in the LDA classifications, likely due to the extensive range of overlapping issues or the variability in user feedback.

GPT-4o-mini

The entropy distributions presented in Figure 11 reveal consistent patterns that reflect the GPT-4o-mini's overall high confidence in its classifications.



(a) KopieID (b) Reisapp (c) MijnOverheid (d) DigiD

Figure 11: Shannon entropy across applications for GPT-4o-mini

For KopieID, the distribution appears bimodal, with the majority of entropy values concentrated around 0.75 and 1.0, and only a few instances extending to higher levels. This concentration of low entropy suggests that GPT-4o-mini confidently assigns topics to reviews in this application, likely due to the relatively straightforward nature of the dataset and the limited diversity of issues encountered.

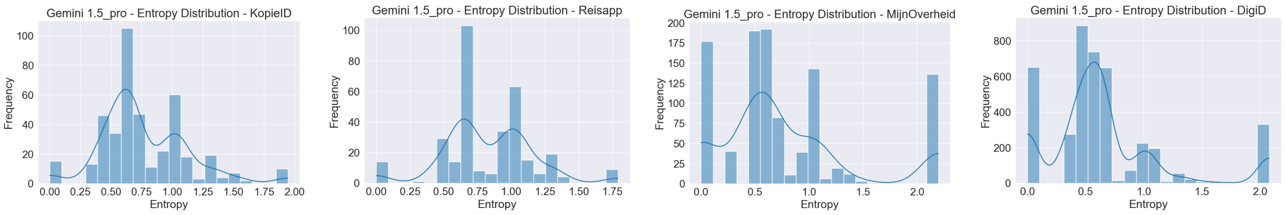
Moving to Reisapp, the distribution remains largely similar, although a slightly broader spread is observed. While the majority of entropy values still fall within lower ranges, there is a minor tail extending towards higher entropy levels. This indicates a modest increase in uncertainty, perhaps due to a slightly more complex range of issues present in Reisapp reviews.

For MijnOverheid, the entropy distribution is generally centered around lower values, approximately 0.5, indicating that the model maintains a reasonable level of confidence despite the potential complexity of the dataset.

Interestingly, DigiD exhibits a pattern somewhat similar to KopieID, with a prominent peak around 0.65. This was notable given that DigiD, the largest dataset in this study, was initially expected to show higher entropy due to the presumed complexity and diversity of its review content, including potentially more nuanced or overlapping topics.

Gemini-1.5-Pro

Turning to Gemini-1.5-Pro, the entropy distributions (see Figure 12) show a clear shift towards higher values compared to GPT-4o-mini, indicating a generally lower level of classification confidence. This is particularly evident for MijnOverheid and DigiD, where the frequency of entropy values exceeding 2 is notably high, suggesting substantial uncertainty in the model's topic assignments for these more complex datasets.



(a) KopieID (b) Reisapp (c) MijnOverheid (d) DigiD

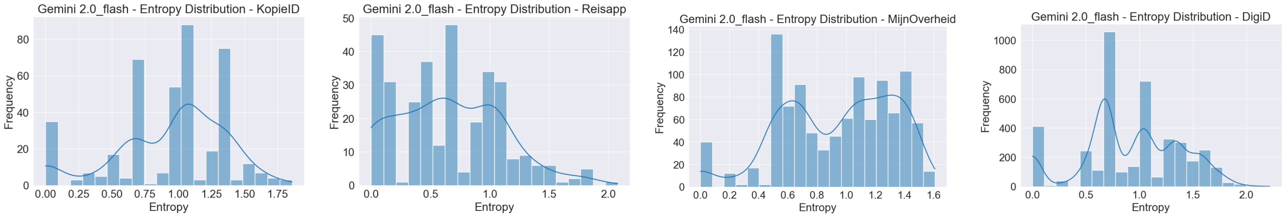
Figure 12: Shannon entropy across applications for Gemini-1.5-Pro

For KopieID, the entropy values are primarily concentrated between 0.75 and 1.0, which is somewhat comparable to the distribution observed for GPT-4o-mini. A similar pattern appears for Reisapp, with the entropy peaking around 1.0, indicating similar certainty in classifications for this application relative to GPT-4o-mini.

For MijnOverheid and DigiD, the entropy values are generally centered around 0.75, initially suggesting a reasonable level of confidence. However, the distributions also reveal a substantial number of instances with very high uncertainty, with entropy values exceeding 2. This indicates that while the model demonstrates some degree of confidence on average, it frequently encounters significant ambiguity in classifying reviews within these more complex datasets.

Gemini-2.0-Flash

Moving on to Gemini-2.0-Flash, the entropy distributions presented in Figure 13 suggest a generally moderate level of confidence, with some variation in more complex datasets.



(a) KopieID (b) Reisapp (c) MijnOverheid (d) DigiD

Figure 13: Shannon entropy across applications for Gemini-2.0-Flash

For KopieID, the distribution appears somewhat bimodal, with a primary cluster around 0.6–0.8 and a secondary group extending up to about 1-1.25. This implies that, while the model is fairly decisive for a large share of KopieID reviews, there remains a subset of cases for which the classification is less certain.

In Reisapp, the distribution is more compact, with the majority of entropy values clustering below 1.0, suggesting improved confidence relative to Gemini-1.5-Pro and GPT-4o-mini.

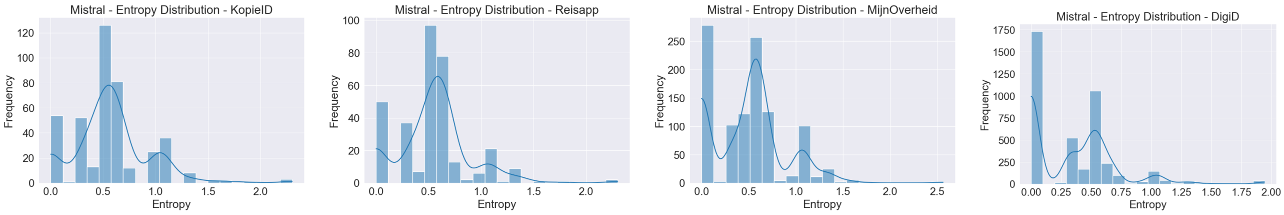
For MijnOverheid, the entropy distribution centers around 0.5–0.7, suggesting a relatively stable level of confidence for many reviews. However, there is a notable spread extending beyond 1.2, highlighting that the model encounters higher uncertainty for a non-trivial fraction of the dataset, possibly due to overlapping or more nuanced issue categories.

Finally, DigiD, the largest dataset in this study, exhibits a multi-modal entropy distribution, with prominent clusters around 0.75 and 1.0, and some values exceeding 1.5. This pattern suggests both a robust core of confidently classified reviews and a subset of higher-uncertainty cases, likely reflecting the complexity and diversity of issues present in DigiD reviews.

Mistral-Large-2411

Turning to Mistral-Large-2411, the entropy distributions across all four applications suggest a consistently lower level of uncertainty compared to the previous models (see Figure 14). For KopieID, the distribution is sharply peaked at lower entropy values, indicating that the model confidently assigns topics for most reviews.

A similar pattern emerges for Reisapp, with a concentration around small entropy values and only a few instances extending to higher ranges. The distributions for MijnOverheid and DigiD—despite their broader sets of issues—also exhibit smaller spreads than those observed for other models, reflecting a robust capacity to handle overlapping or nuanced topics.



(a) KopieID (b) Reisapp (c) MijnOverheid (d) DigiD

Figure 14: Shannon entropy across applications for Mistral-Large-2411

Observation 4:

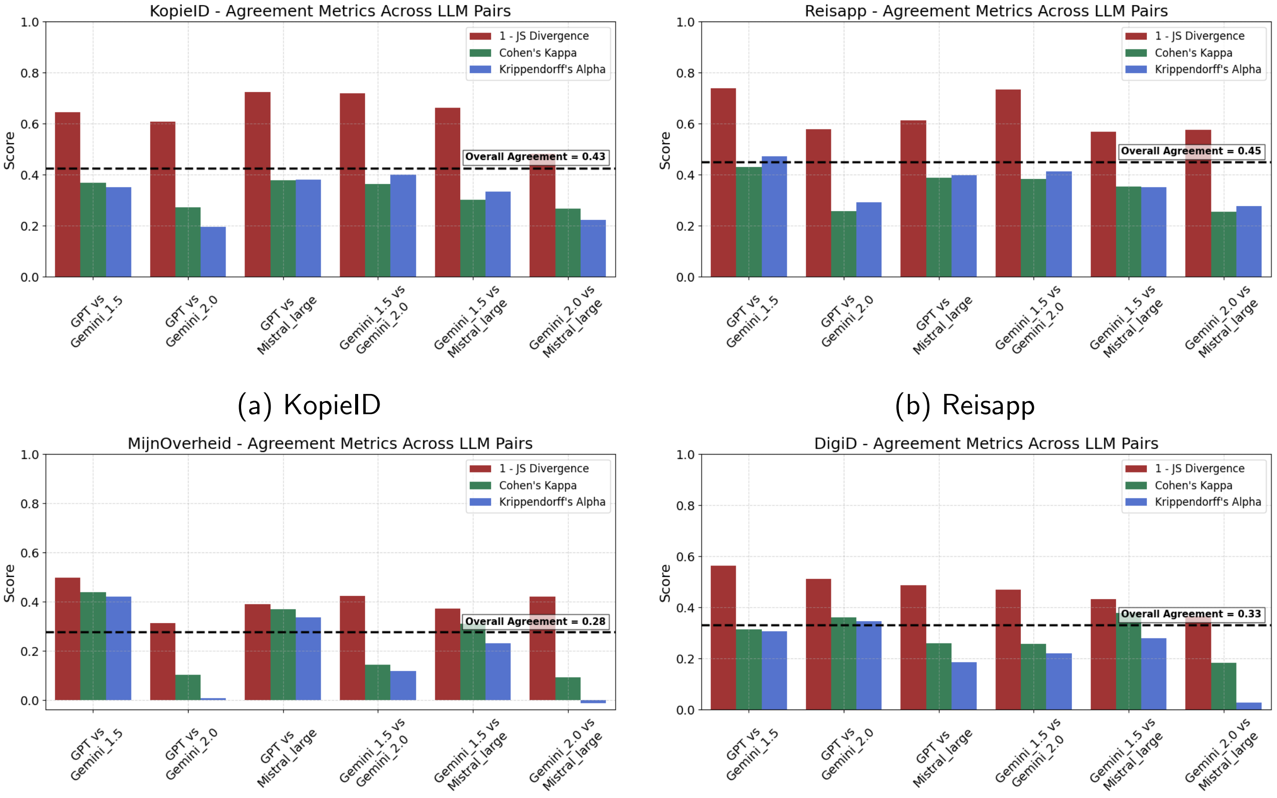
Although Mistral-Large-2411 demonstrated strong confidence in most applications, its performance on DigiD was less consistent. Multiple runs were required due to unpredictable outputs, and several newly introduced topics had to be removed during final classification. These practical challenges highlight the importance of reliability and stability in evaluating model performance.

> Taking into account both the Shannon entropy and the coherence scores in Table 5, as well as the reliability issues encountered with Mistral-Large-2411, Gemini-2.0-Flash was considered one of the best-performing models. Accordingly, it has been selected for subsequent analyses to ensure robust and coherent results.

5.2.2 Agreement

In this section, the consistency among the classification outputs is evaluated using three agreement metrics: Jensen–Shannon Divergence, Cohen's Kappa, and Krippendorff's Alpha. Because lower JS Divergence indicates greater alignment among probability distributions, $1 - JS$ is presented to maintain a consistent "higher-is-better" scale with the other metrics. For each plot from Figure 15, the overall agreement across all metrics and model pairs has been computed and is represented by a bold dashed line. These metrics are essential for assessing the reliability of multi-label classifications, with higher agreement reflecting stronger model alignment and greater confidence.

The agreement metrics across LLM pairs reveal several meaningful patterns regarding the consistency of issue classifications. Notably, the $1 - JS$ Divergence consistently shows higher scores than the other two metrics across all applications. This outcome is expected, given that $1 - JS$ Divergence assesses the similarity between probability distributions (in this case, the relevance scores assigned by the models to each issue). Since it operates on continuous data rather than categorical labels, it naturally produces higher values, reflecting similarity in the distribution of predictions rather than exact categorical matches.



(a) KopieID

(b) Reisapp

(c) MijnOverheid

(d) DigiD

Figure 15: Agreement metrics across LLM pairs for each application. The bold dashed black line represents the overall average agreement across all metrics and model pairs.

In contrast, Cohen's Kappa and Krippendorff's Alpha, while generally lower than the 1-JS Divergence scores, capture agreement in a stricter way than probabilistic similarity measures. Cohen's Kappa remains based on categorical alignment, requiring exact matches between predicted labels to register agreement. However, Krippendorff's Alpha, computed here using an interval scale, accounts for the degree of difference between predictions, meaning that near matches are penalized less severely than complete mismatches. Despite this more flexible evaluation, the overall agreement remains moderate, reflecting substantial variation in how the models assign relevance scores and categorize issues. A particularly striking example appears in the MijnOverheid dataset, where Krippendorff's Alpha drops below zero for the Gemini-2.0-Flash and Mistral-Large-2411 pair, indicating systematic disagreement (worse than random chance). This suggests that these models likely diverged in their interpretation of the issues in MijnOverheid, possibly because of the dataset's broader and more general issue categories, which leave more room for subjective model interpretations.

When looking at the overall agreement scores, represented by the dashed black lines, values range between 0.28 for MijnOverheid and 0.45 for Reisapp. This indicates a fair or moderate level of alignment at best between the LLMs across applications.

These results illustrate that, although LLMs recognize similar overarching themes, their specific issue categorizations often differ, especially in larger datasets, like MijnOverheid or DigiD. Moreover, part of this variability likely stems from architectural differences between the models and their varying sensitivity to linguistic nuances or context. Some models may emphasize technical terminology, while others may capture user sentiment or contextual phrasing more strongly.

> In summary, the agreement analysis suggests that while probabilistic agreement (1-JS Divergence) indicates reasonable alignment between models, categorical agreement remains more challenging. The results highlight the challenges of achieving high consistency in issue classification, especially in complex multi-label environments such as app review analysis.

5.3 Issue-Star Rating Assessment

Following the classification procedure, two additional analyses were carried out to assess the extracted issues in relation to the app ratings. The first analysis evaluated the overall effect of each issue on the star ratings by applying CLMs, providing insights into how the presence and intensity of specific issues influence user satisfaction. The second analysis examined the evolution of issue effects on star ratings over time, revealing whether certain issues have become more or less influential in recent years. As previously established, only the best-performing model was used for these analyses. Based on the earlier evaluation of classification performance, Gemini-2.0-Flash was selected as the model for this part of the study. This ensured that the findings are based on the most reliable classification outputs obtained.

These analyses aim to provide a deeper understanding of the practical relevance of the identified issues, both in terms of their immediate association with user satisfaction and their progression over time.

5.3.1 Effect Analysis

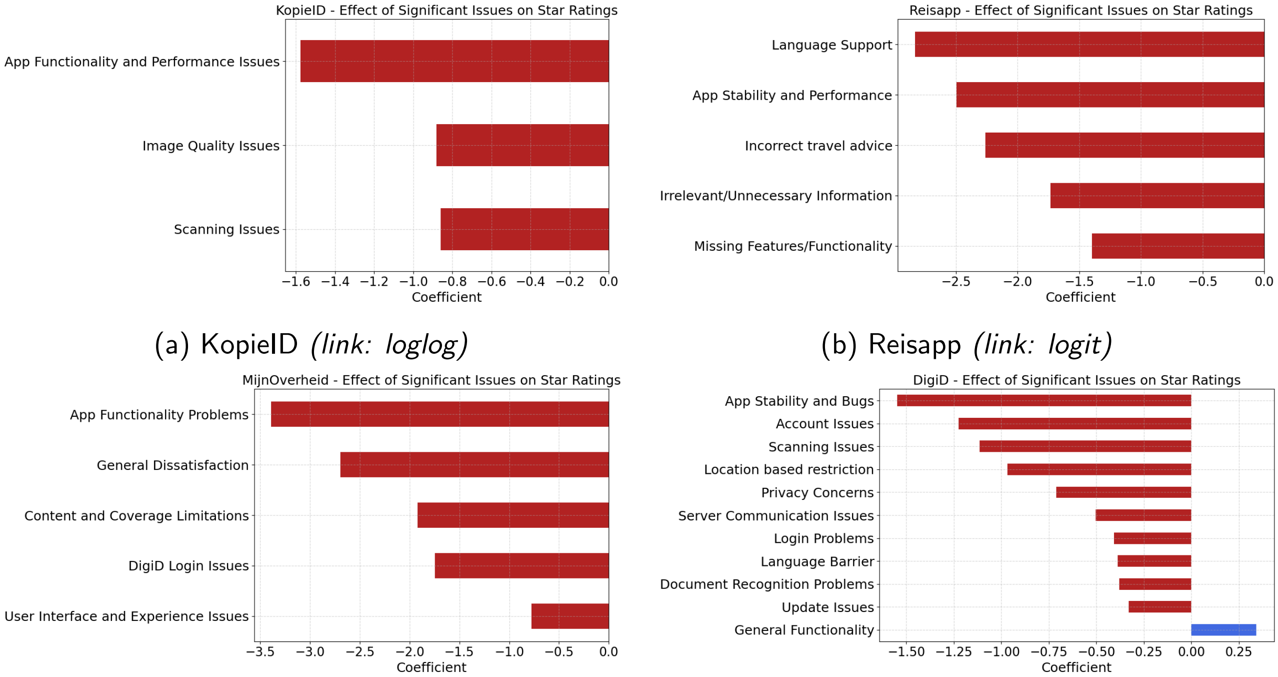
Before the results are presented, the expectations for this analysis should be outlined. Generally, issues reported in user reviews are anticipated to be significant and to correlate negatively with app star ratings, as they typically reflect user dissatisfaction or functional shortcomings. In this section, it is examined whether this assumption holds true across different applications.

The results are presented in Figure 16, each subfigure corresponding to one applications: KopieID, Reisapp, MijnOverheid, and DigiD. In these plots, the x-axis represents the estimated coefficients from CLMs, indicating the direction and strength of each issue's effect on the star ratings. Negative coefficients reflect issues associated with lower ratings, whereas positive coefficients correspond to topics linked to higher user satisfaction. Each bar represents a significant issue according to the CLM analysis, where significance is defined by a p-value not greater than 0.05. This provides a clear visual hierarchy of the issues most strongly influencing app ratings. Notably, the number of significant issues is less than or equal to the total number of issues identified in the dataset. For a more detailed view of the regression results, including the exact coefficient estimates, standard errors, z-values, and p-values, the full tables are provided in Appendix A.

The visualizations show a clear overview of how different issues influence user star ratings across the four government applications. In all cases, the majority of issues exhibit a negative impact on ratings, which aligns well with the initial expectations: user-reported issues typically signal dissatisfaction, especially when they concern functionality, technical stability, or core services of the application.

For example, in Reisapp and MijnOverheid, very strong negative effects can be observed for critical operational issues."Language Support", "App Stability and Performance", and " Incorrect travel advice" in Reisapp, as well as "App Functionality Problems" and "General Dissatisfaction" in MijnOverheid, are among the most impactful. This underlines that when users encounter barriers to basic accessibility or experience malfunctioning features, they tend to penalize the app heavily in their reviews. Similarly, DigiD and KopieID show pronounced negative coefficients for essential functionalities: "App Stability and Bugs", "Account Issues" and "Scanning Issues", are leading factors in user dissatisfaction for DigiD, whereas “App Functionality and Performance Issues” dominates the negative side for KopieID.

Interestingly, however, one issue shows a positie effct on the star rating, namely DigiD’s “General Functionality”, opening a nuanced perspective. This might suggest that in some cases, users not only report problems but also provide constructive feedback or positive recognition for well-functioning aspects of the app. It is possible that users, even when pointing out minor issues, take the opportunity to acknowledge features that they appreciate or find useful. Another possible explanation is that some reviews may include improvement suggestions phrased in a neutral or even slightly positive tone.



(a) KopieID *(link: loglog)*

(b) Reisapp *(link: logit)*

(c) MijnOverheid *(link: logit)*

(d) DigiD *(link: loglog)*

Figure 16: The estimated effects of each issue on star ratings, as computed using CLMs, are presented for each application, and selected according to Chapter 3.

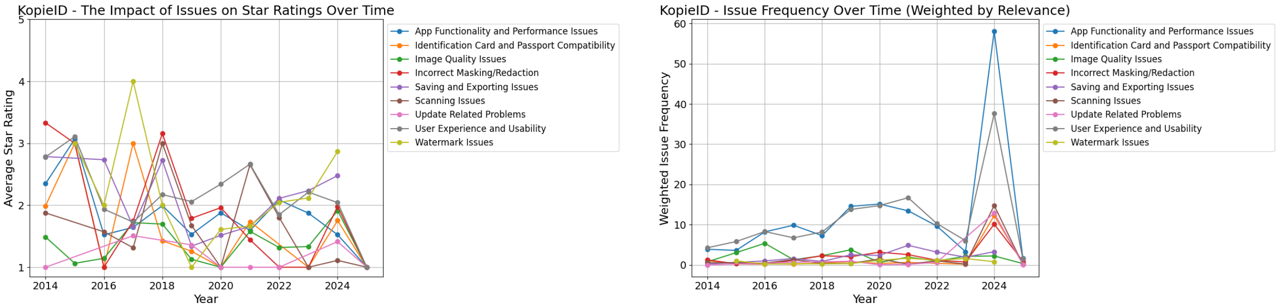
> Overall, the results highlight that technical reliability and core functionalities remain the primary drivers of user dissatisfaction, while aspects related to the general functionality can sometimes positively influence star ratings. This suggests that users appreciate and acknowledge well-executed features, especially when they provide constructive feedback or balance their criticism with recognition of the app’s strengths.

5.3.2 Time Analysis

In this section, the evolution of the impact of issues identified and classified by Gemini-2.0-Flash is examined in relation to the star ratings. Building upon the previous classification results, this temporal analysis aims to provide insights into whether the influence of certain issues has intensified, diminished, or remained stable across different periods. The constructed timeline graphs incorporate LLM-assigned relevance scores, allowing issue frequency to be weighted by estimated importance in each review. In this way, both occurrence and perceived significance are accounted for, fully leveraging Gemini-2.0-Flash’s classification capabilities.

KopieID

The temporal analysis of KopieID, as shown in Figure 17, provides a comprehensive view of how the identified issues evolved in both their frequency and their impact on user satisfaction over time.



(a) Evolution of the impact of identified issues on star ratings over time, based on classifications provided by Gemini-2.0-Flash.

(b) Weighted frequency of user-reported issues over time, based on relevance scores assigned by Gemini-2.0-Flash.

Figure 17: Temporal analysis for KopieID

The left-hand plot illustrates the evolving relationship between reported issues and the average star ratings they received. Here, lower star ratings for an issue indicate a greater negative impact on user satisfaction. Notably, "Update Related Problems" and "Image Quality Issues" consistently exhibit the strongest negative effect across the years.

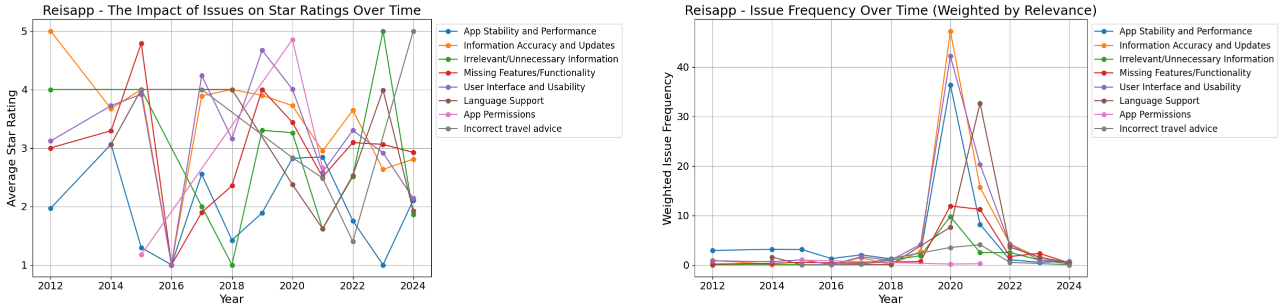
Similarly, "Scanning Issues" and "App Functionality and Performance Issues" display fluctuating yet consistently negative impacts, particularly noticeable during the earlier years of the timeline. Additionally, "Incorrect Masking/Redaction", which appeared to be a relatively minor concern in the initial stages of the application's lifecycle, gradually emerged as a more significant driver of user dissatisfaction in later years. These patterns highlight users' pronounced sensitivity to technical and operational shortcomings, with such issues leading to harsher penalization of the app when they become more prominent or persistent.

An interesting observation emerges regarding "Watermark Issues", which consistently received ratings around 4 stars, making it the only issue associated with relatively high user satisfaction. At first glance, this might suggest a surprisingly positive perception of this issue. However, when this finding is correlated with the weighted frequency plot, it becomes clear that the occurrence of "Watermark Issues" is minimal, remaining close to zero throughout the entire timeline. This low frequency indicates that, although the impact appears high in isolation, its overall significance in shaping user satisfaction is negligible.

Looking more closely at the weighted frequency plot (right-hand plot), there is a significant spike in the frequency of both "App Functionality and Performance Issues" and "User Experience and Usability" around 2023–2024, which corresponds closely with the sharp decline in star ratings for the former in the impact plot. This alignment suggests a direct link between the rise in functional problems and increasing user dissatisfaction. However, despite the heightened frequency of "User Experience and Usability" concerns, its impact on star ratings remained moderate, fluctuating between 2 and 3 stars throughout the timeline. This pattern implies that while usability issues became more commonly reported, they did not evoke the same intensity of negative sentiment, possibly reflecting users' partial satisfaction with the app's design and overall user interface.

Reisapp

The temporal analysis of Reisapp, shown in Figure 18, reveals highly dynamic patterns in both the impact of reported issues on star ratings and their weighted frequency over time. Compared to KopieID, Reisapp displays a much more erratic trajectory, especially in the earlier years of the timeline.



(a) Evolution of the impact of identified issues on star ratings over time, based on classifications provided by Gemini-2.0-Flash.

(b) Weighted frequency of user-reported issues over time, based on relevance scores assigned by Gemini-2.0-Flash.

Figure 18: Temporal analysis for Reisapp

Between 2012 and 2018, the dataset appears relatively sparse, with most issues occurring fewer than 10 times per year — a low frequency that likely contributes to the wide fluctuations observed in the impact scores. During this period, issues like "Incorrect travel advice" and "Missing Features/Functionality" swing between low and high impact, with some issues surprisingly reaching very high ratings (between 4 and 5 stars), and others dropping to extremely low ratings around 2016. This volatility suggests that with a limited number of reviews, individual experiences disproportionately influenced the overall trend.

A clearer, more interpretable pattern emerges after 2018. Starting in 2019, there is a notable surge in the frequency of reported issues, peaking sharply in 2020 — coinciding with the global outbreak of the SARS-CoV-2 pandemic. This spike in issue reporting aligns closely with a visible decline in star ratings across most issues in the same period. Particularly, technical and accessibility-related concerns such as "App Permissions", "Language Support", and "App Stability and Performance" became more prominent in user feedback. The heightened reliance on digital tools for travel information during the pandemic likely amplified user sensitivity to such problems, as expectations for timely updates and reliable functionality increased dramatically.

Following the 2020 peak, both the frequency of reported issues and their negative impact on star ratings gradually declined. By 2022, the data indicates a partial recovery, as the sharpest dips in ratings begin to stabilize. This suggests that as pandemic pressures eased and app usage patterns normalized, user expectations may have adjusted, and developers might have addressed the most critical issues that surfaced during the earlier spike.

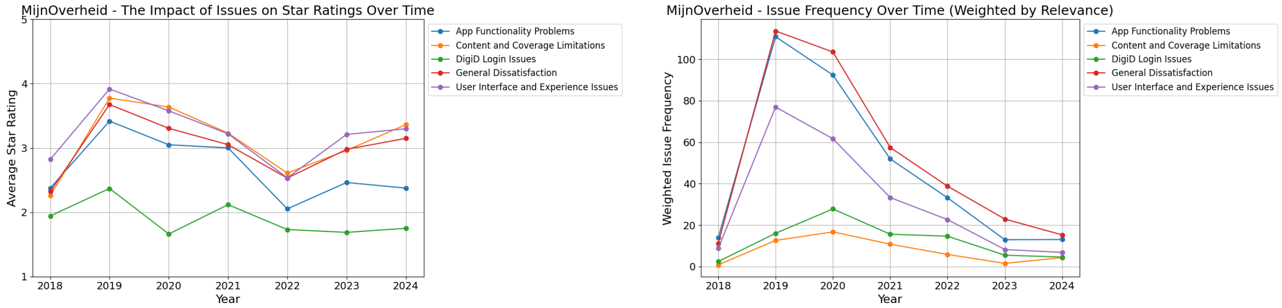
MijnOverheid

The temporal analysis of MijnOverheid, shown in Figure 19, presents a more structured picture compared to previous applications, indicating clearer trends in both issue frequency and their impact on user ratings over time.

Starting with the issue frequency (right-hand plot), a sharp increase in the reported problems is observed in 2019. "General Dissatisfaction" and "App Functionality Problems" are particularly prominent, both showing significant surges in weighted frequency. Notably, User "Interface and Experience Issues" also follow a similar trajectory, albeit at a lower level of occurrence, suggesting that concerns about the app's usability, while present, were not the primary driver of user attention during this period.

This rise in reported issues corresponds closely with a parallel decline in star ratings observed in the impact plot (left-hand side). In the years leading up to 2020, issues such as "App Functionality

Problems", "General Dissatisfaction", and "Content and Coverage Limitations" all showed a deterioration in their associated star ratings, reflecting growing user frustration. Interestingly, "DigiD Login Issues" — a category specific to interactions between MijnOverheid and related services — maintains a consistently negative impact on ratings, with little recovery throughout the observed period. This persistent dissatisfaction suggests that integration or interoperability problems remained unresolved, continuing to erode user satisfaction.



(a) Evolution of the impact of identified issues on star ratings over time, based on classifications provided by Gemini-2.0-Flash.

(b) Weighted frequency of user-reported issues over time, based on relevance scores assigned by Gemini-2.0-Flash.

Figure 19: Temporal analysis for MijnOverheid

Notably, across all five identified issues, the peak of user dissatisfaction consistently occurs around 2022, indicating a period of widespread user frustration that cut across functional, usability, and integration concerns. Although the frequency of reported issues had already begun to decline by this point, the lingering negative perception likely reflects cumulative dissatisfaction built up over preceding years.

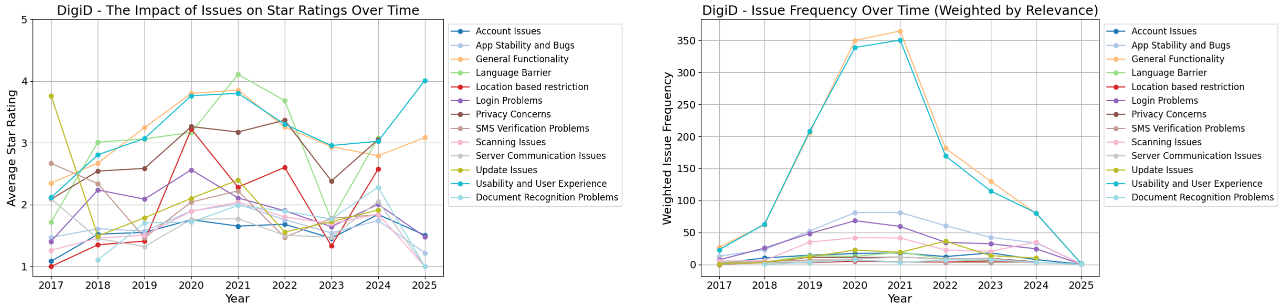
The relatively structured nature of these trends, compared to the volatility seen in other applications, suggests that MijnOverheid's challenges were more systemic and recognized by users over time, rather than episodic spikes in dissatisfaction. This is likely influenced by the fact that MijnOverheid's dataset is almost three times larger than those of the previous two applications, providing a more stable and representative view of user concerns and helping to smooth out irregular fluctuations observed in smaller datasets.

DigiD

Figure 20 reveals the temporal analysis for the fourth application, namely DigiD. With a larger dataset and broader user base, the trends observed here are again interpretable, providing meaningful insights into how issues evolved over time and how they impacted user satisfaction.

Starting with the weighted frequency plot (right-hand side), a sharp and distinctive rise in reported issues is observed beginning in 2018 and peaking decisively around 2021. In particular, "General Functionality" and "Usability and User Experience" dominate the landscape, each reaching peak frequencies of approximately 350 weighted occurrences, significantly surpassing all other issue categories. However, these categories represent broader user perceptions and do not pinpoint specific technical failures; instead, they capture general dissatisfaction with the app's overall performance and design. Given their generic nature, it becomes essential to shift the focus toward more technically specific issues, such as "App Stability and Bugs", "Login Problems", and "Scanning Issues". These categories clarify the operational challenges users faced and help explain how specific technical issues influence user satisfaction over time.

Turning to the impact plot (left-hand side), the aforementioned technical issues are seen to align with some of the lowest average star ratings, clearly reflecting heightened user dissatisfaction. This correlation between increasing issue frequency and declining star ratings highlights a strong interplay between user experience and technical performance for DigiD, particularly during the period from 2021 to 2024. Notably, the year 2021 marks the lowest point in star ratings across nearly all categories, indicating a peak in user frustration during this time.



(a) Evolution of the impact of identified issues on star ratings over time, based on classifications provided by Gemini-2.0-Flash.

(b) Weighted frequency of user-reported issues over time, based on relevance scores assigned by Gemini-2.0-Flash.

Figure 20: Temporal analysis for DigiD

Interestingly, after 2021, both the frequency of reported issues and their negative impact on star ratings show signs of improvement. By 2023–2024, issue frequencies decline sharply, and star ratings begin to recover, especially for previously critical issues like “Language Barrier” and “Location based restriction”.

The generics: “General Functionality” and “Usability and User Experience” stand out with a more stable or even slightly positive trend in ratings throughout the entire timeline, regardless of issue frequency. This stability implies that users continued to appreciate core functionalities and design usability, even during periods of heightened technical issues.

DigiD’s temporal analysis reveals a strong, responsive relationship between technical issues and user satisfaction trends. The clear spike and subsequent decline in issue frequency, mirrored by fluctuations in user ratings, illustrate how critical it is for public services to maintain consistent technical reliability.

> The temporal analysis highlights that clearer patterns emerge in larger datasets, revealing that the most significant declines in user satisfaction occurred between 2020 and 2023, largely driven by technical issues. Interestingly, broader, non-technical concerns such as usability and general functionality tend to be associated with higher ratings, suggesting that users continued to appreciate well-designed aspects of the apps even amidst technical difficulties.

5.4 Forecasting

In the final stage of this analysis, a forecasting exercise was performed to anticipate potential future issues in app performance and user satisfaction. This task was conducted exclusively using the Gemini-2.0-Flash LLM, selected as the best-performing model in earlier stages of the study. The model leveraged historical app reviews — incorporating the content of the reviews, their timestamps, and the distribution of identified issues — to detect emerging patterns and predict possible issues that could arise after 2023.

The historical dataset, comprising all reviews up to and including 2023, was incorporated directly into the prompt to inform the model’s predictions, while the reviews from 2024 and 2025, representing new and unseen data, were used to evaluate the model's output. A summary table detailing the number of reviews considered as historical input for each application is provided in Table 6.

	KopieID	Reisapp	MijnOverheid	DigiD
historical data	259	322	1007	3838
new data	157	3	45	299

Table 6: Table of considered historical data (until and including reviews from 2023). The new data refers to all the reviews from 2024 and 2025.

For each predicted issue, Gemini-2.0-Flash estimated the probability of occurrence, offering insights into the likelihood of future challenges. In addition to forecasting potential issues, the model also provided recommendations aimed at helping app owners proactively prevent or mitigate these problems.

For each application, a dedicated analysis is presented comparing the forecasted issues with the actual issues observed in the 2024–2025 dataset. These comparisons are visualized in the plots 21 - 24. Each label in the plots follows the format: *forecasted_issue* $\sim$ *actual_issue* $\sim$ *similarity_score*, where the similarity score is computed using the *all-MiniLM-L6-v2* sentence transformer model. The visual contours indicate the quality of the match: predicted issues with a similarity score of 0.6 or higher are highlighted in green, those below 0.6 are marked in red, and actual issues that were not matched with any forecasted label are contoured in gray. These visualizations provide a clear assessment of the forecasting accuracy in both issue identification and frequency estimation.

KopieID

The forecasting results for KopieID from Figure 21 reveal a mixed performance in identifying future issues.

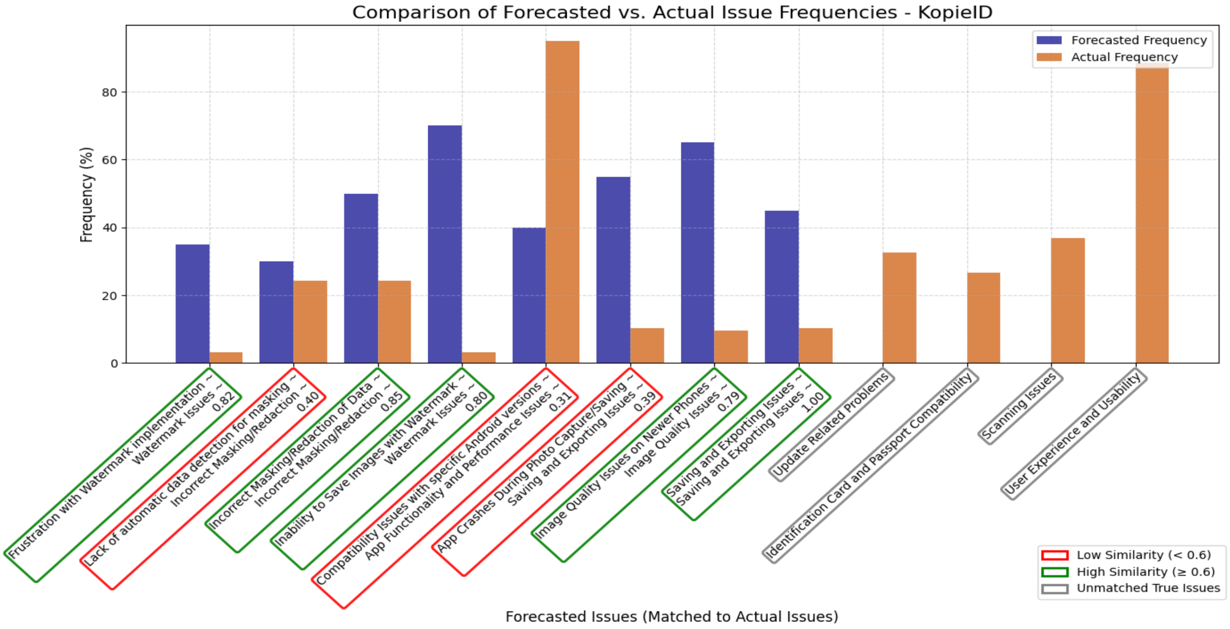


Figure 21: Evaluation of the forecasted issues for KopieID and their predicted frequencies, conducted by comparing them against the true labels assigned during the classification process.

Several forecasted issues achieved a high semantic similarity (≥ 0.6) with the actual issues from 2024–2025, as indicated by the green contours in the plot. Notably, the model successfully anticipated problems related to “Incorrect Masking/Redaction”, “Watermark Issues”, “Saving and

Exporting Issues", and "Image Quality Issues". These high-similarity matches suggest that the model effectively captured recurring technical challenges based on historical patterns.

However, the presence of several forecasted issues with lower similarity scores (highlighted in red) points to limitations in the model's predictive precision. Issues such as "Compatibility Issues with Specific Android Versions", "App Crashes During Photo Capture", and "Lack of automatic detection for masking" were identified by the model but did not align closely with the actual issues observed, indicating potential noise or over-specification in the forecasting output.

Moreover, the plot highlights unmatched true issues (contoured in gray), such as "Scanning Issues" and "User Experience and Usability." These were not predicted by the model despite being high-impact based on issue frequency, and, in the case of "Scanning Issues," also according to the CLM analysis (see Figure 16). Additionally, the model tends to overestimate the prevalence of certain forecasted issues compared to their actual occurrence. For example, "Incorrect Masking/Redaction" and "Watermark Issues" were forecasted at notably higher frequencies than observed in the real data.

Alongside the forecasting results, the LLM provided a set of recommendations for KopieID, focusing on both technical improvements and user experience enhancements. These suggestions address key forecasted issues, such as compatibility testing for new Android devices, improving image capture quality, and strengthening saving and exporting functionalities. Notably, the recommendations also include proactive, user-focused measures like adding biometric security and offering more flexible image-saving options.

LLM Suggestions - KopieID

- Implement robust testing on new Android devices upon release to prevent compatibility issues.
- Add the ability to save images with the watermark.
- Improve image capture quality by using the native camera functionality or updating the camera API.
- Implement automatic detection of sensitive data for redaction.
- Conduct thorough testing for saving and exporting functionalities on different devices and platforms.
- Offer more granular control over watermark placement and appearance.
- Enhance the UI to allow adjustments on the masking after creation.
- Implement better error handling and user feedback for saving and exporting processes.
- Regularly update the app to maintain compatibility with newer devices and Android versions.
- Provide the option to save images as JPEG or PNG formats in addition to PDF.
- Add a password or biometric security feature to prevent unauthorized access to saved images within the app.

Reisapp

The forecasting results for Reisapp presented in Figure 22 show a moderate alignment between predicted and actual issues, with a few notable successes. The LLM accurately anticipated several key concerns, such as "App Stability and Performance", "Information Accuracy and Updates" and "Language Support", which are marked in green.

However, the analysis also reveals clear gaps. Several high-frequency actual issues, including "Incorrect travel advice", "User Interface and Usability", and "Missing Features/Functionality", remained unmatched in the forecast, as indicated by the gray contours. These omissions suggest that the model struggled to anticipate some of the more application-specific or evolving concerns.

On the forecasted side, some predicted issues like "Difficulty Navigating the App" and "UI Issues with Dark Mode" fell below the similarity threshold (contoured in red), highlighting instances where the model's predictions were poorly aligned with the actual user feedback from 2024–2025.

When examining the frequencies, the model tended to underestimate the prominence of certain high-impact issues, particularly "App Stability and Performance" and "Information Accuracy and

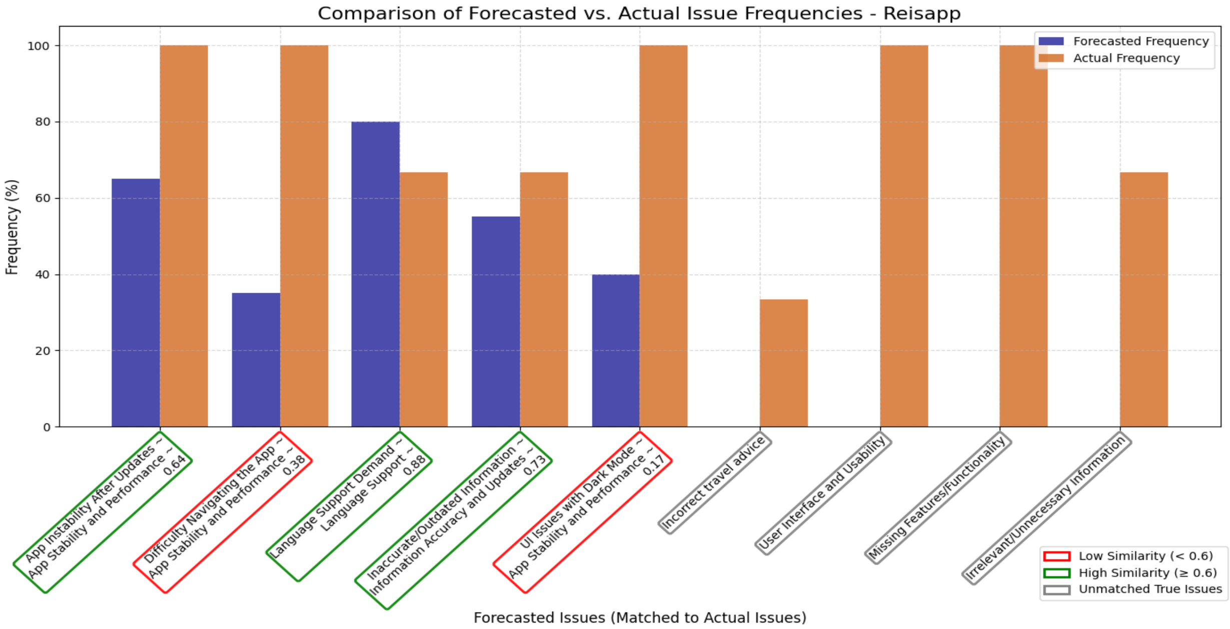


Figure 22: Evaluation of the forecasted issues for Reisapp and their predicted frequencies, conducted by comparing them against the true labels assigned during the classification process.

Updates". In contrast, for "Language Support", the model slightly overestimated the expected frequency compared to the actual occurrence.

Complementing the forecasting results, the LLM provided recommendations for Reisapp. Key suggestions include enhancing multi-language support, improving the accuracy of travel information through real-time data verification, and addressing UI issues, particularly for dark mode compatibility. Additionally, the recommendations emphasize performance optimization, clearer communication through push notifications, and refining the app's navigation to better guide different user groups.

LLM Suggestions - Reisapp

- Prioritize adding multi-language support, especially English, considering the number of international users and expats in the Netherlands.
- Implement a more robust testing process for app updates to prevent crashes and loading issues post-release. Consider beta testing with a subset of users.
- Improve the update frequency and accuracy of travel information, potentially by integrating with real-time data sources and verifying information with local sources.
- Address UI issues, particularly related to dark mode, to ensure text and elements are readable in all themes. Conduct thorough testing on different devices and screen settings.
- Revamp the app's navigation to make it more intuitive, possibly by simplifying the information architecture and improving the search functionality. Conduct user testing to identify pain points.
- Implement better error handling and user feedback mechanisms to help users resolve issues and provide valuable diagnostic information to the development team.
- Ensure the app clearly distinguishes between requirements for different user groups (e.g., residents vs. tourists) to avoid confusion.
- Optimize the app's performance to reduce loading times and improve responsiveness, especially after push notifications.
- Regularly audit and remove irrelevant or outdated information to streamline the user experience.
- Enhance the clarity of push notifications to provide specific details about changes in travel advisories, rather than just alerting users to a change.

MijnOverheid

Figure 23 shows a relatively poor alignment between the predicted and actual issues for MijnOverheid, though with a few notable forecasted issues.

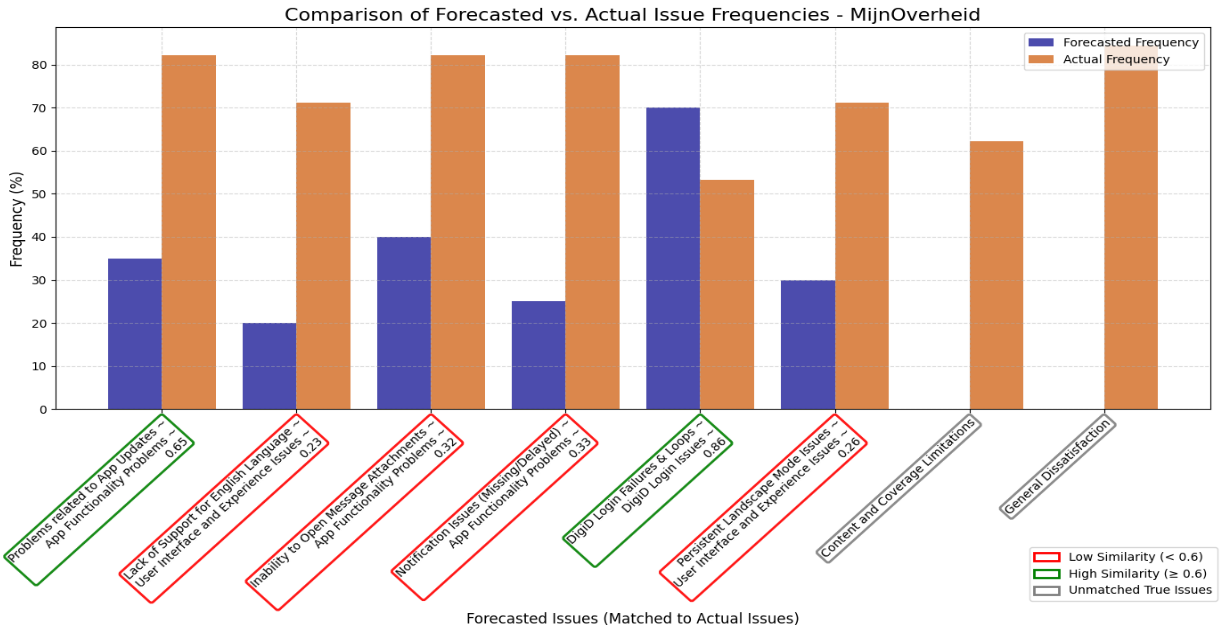


Figure 23: Evaluation of the forecasted issues for MijnOverheid and their predicted frequencies, conducted by comparing them against the true labels assigned during the classification process.

Encouragingly, the LLM successfully anticipated two critical areas: "App Functionality Problems" and "DigiD Login Issues". These accurate forecasts indicate that the model effectively captured some of the recurring, systemic problems that were also prominent in the historical data. Moreover, the predicted frequency of the former issue was reasonably close to the actual one.

However, the model also produced lower similarity matches, particularly for one of the broader categories, namely "User Interface and Experience Issues". The plot further highlights a limitation: important actual issues such as "Content and Coverage Limitations" and "General Dissatisfaction" remained unmatched in the forecast.

Although the similarity scores between the following forecasted issues and the actual ones remain relatively low, the predicted topics are still noteworthy: "Lack of Support for English Language," "Inability to Open Message Attachments," "Notification Issues (Missing/Delayed)," and "Persistent Landscape Mode Issues." These represent more specific and technical concerns, which could reasonably be interpreted as subcategories or concrete manifestations of the broader actual issues.

The LLM-generated recommendations for MijnOverheid effectively address both forecasted and actual challenges. They prioritize DigiD integration and login loop fixes (key predicted issues) and suggest technical improvements like app compatibility, landscape mode support, and crash prevention. Additionally, they tackle lower-similarity forecasted issues such as notification failures and limited language support, which still impact user experience.

LLM Suggestions - MijnOverheid

- Improve DigiD integration and address login loops by thoroughly testing the interaction between the MijnOverheid app and the DigiD app, especially after updates to either app.
- Implement robust error handling and provide clear, user-friendly error messages to guide users when issues occur.
- Address landscape mode issues by ensuring the app respects device rotation settings or provides an in-app setting to lock the orientation.

- Optimize app performance to reduce loading times and prevent crashes, especially when opening message attachments.
- Implement proper notification settings to prevent missing or delayed notifications. Provide an option to customize notification preferences.
- Add multi-language support, particularly English, to cater to a wider user base and improve accessibility.
- Enhance testing procedures to identify and resolve issues before releasing app updates to prevent disruptions.
- Improve app compatibility with a wide range of Android devices and versions.
- Streamline the process of setting up a new pin, providing a clear explanation and guidance to users on how to set up a pin if needed.

DigiD

The forecasting results presented in Figure 24 reveal a notable gap between forecasted and actual issue frequencies in DigiD's case, with most of the actual issues remaining unmatched (gray labels), while the forecasted issues frequently show low similarity scores (red labels). This indicates that the model struggled to anticipate the specific types of issues users faced in 2024–2025.

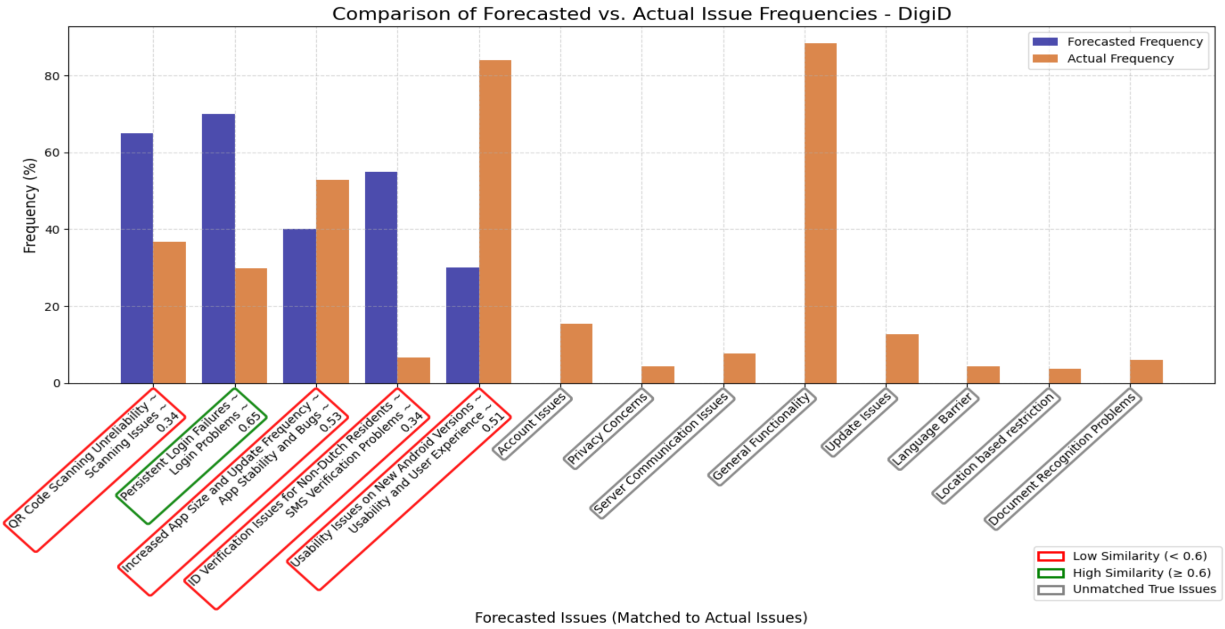


Figure 24: Evaluation of the forecasted issues for DigiD and their predicted frequencies, conducted by comparing them against the true labels assigned during the classification process.

The LLM correctly predicted "Login Problems", being a recurring technical challenge in the historical data. However, the forecasted frequency is notably higher than the actual occurrence, suggesting an overestimation of the issue's future relevance.

In contrast, the most impactful issues in the actual data, such as "General Functionality", "Usability and User Experience", and "App Stability and Bugs", were either associated with lower similarity scores or were not captured at all in the forecast. These omissions are particularly critical, given that the first two issues represented the highest actual frequencies in the dataset. This discrepancy suggests that the model may have disproportionately focused on more specific, narrowly defined issues (such as: "ID Verification Issues for Non-Dutch Residents"), overlooking broader, general categories that proved to be highly relevant in the users' feedback.

Gemini's recommendations for DigiD focus on strengthening technical reliability and user accessibility, directly addressing critical areas such as login failures, QR code scanning, and cross-device compatibility. It also emphasizes performance optimization, through testing, and expanded login options like SMS and biometrics. Notably, simplifying the user interface and improving language support could enhance overall usability for citizens.

LLM Suggestions - DigiD

- Implement more robust error handling and logging to quickly identify and address the root causes of login failures and QR code scanning issues.
- Improve QR code scanning reliability by optimizing camera integration and ensuring compatibility across various devices and screen settings (e.g., dark mode).
- Provide clear and actionable error messages to guide users in troubleshooting common problems.
- Offer alternative login methods (e.g., SMS verification, backup codes) to ensure users can access their accounts even when the app is malfunctioning.
- Streamline the ID verification process to support a wider range of identification documents, including those from non-Dutch residents and various EU countries.
- Optimize the app's size and reduce update frequency to minimize storage space and data usage concerns.
- Thoroughly test app updates on a variety of devices and Android/iOS versions before release to prevent compatibility issues.
- Enhance the user interface (UI) and user experience (UX) to simplify navigation and reduce the number of steps required for common tasks.
- Prioritize accessibility and language support (e.g., offering an English language option) to cater to a diverse user base.
- Implement biometric authentication (e.g., fingerprint, facial recognition) for faster and more secure login.

> The forecasting analysis demonstrated the potential of LLMs to anticipate emerging app issues, successfully identifying several relevant and specific problem areas across applications. However, the model tended to focus predominantly on narrowly defined issues, while overlooking broader concerns. Additionally, its ability to accurately predict the frequency of these issues proved somewhat limited (with a few notable exceptions) which is understandable given the complex and dynamic nature of the datasets. Nonetheless, the generated recommendations were clear, actionable, and well-structured, typically amounting to around ten targeted suggestions per application.

6 Discussion

This chapter discusses the main findings in relation to the three research questions, structured around their theoretical and practical implications. It highlights how language technologies, particularly LLMs, reshape the way user feedback is analyzed in the public sector, offering both a deeper understanding of user behavior and valuable insights for service improvement. Connecting these thesis results to established theories and real-world applications demonstrates how NLP can support more citizen-aligned and responsive digital government services.

6.1 Theoretical Implications

6.1.1 RQ1

One of the theoretical contributions of this study lies in rethinking how user concerns are surfaced and structured through language technologies. A key aim of RQ1 was to assess whether LLMs can extract coherent and meaningful issue categories without relying on predefined taxonomies, as well as to evaluate their ability to accurately classify user reviews based on these categories. This connects directly to the negativity bias and service quality theory. The ability to surface specific, recurring issues from user feedback aligns with how negative experiences are typically reported and highlights the importance of addressing them to improve public service delivery.

The comparative analysis of issue extraction methodologies reveals significant advantages of LLM-based approaches. The quantitative evaluation of coherence scores across all four applications demonstrated that LLMs consistently outperformed traditional LDA methods in producing semantically coherent issue clusters (Table 5). Notably, Gemini-2.0-Flash, Mistral-Large-2411, and Claude-3.5-Sonnet-241022 exhibited superior capabilities in identifying cohesive yet nuanced user concerns, providing a more refined foundation for subsequent analysis.

Methodological convergence was assessed through cosine similarity analysis between LLM-generated issues and LDA topics using vector embeddings of influential terms. The moderate overlap confirmed shared thematic concerns, while divergences highlighted the distinct analytical strengths of each method. Moreover, LLMs effectively group related concerns into clear, labeled categories, enabling faster issue identification. While LDA offers less coherence and no labeling, it can surface more specialized or niche topics that LLMs may subsume into broader categories, highlighting a trade-off between clarity and granularity.

The cluster analysis of LLM-extracted issues further supports their effectiveness, showing consistent identification of key user concerns across different models. In the context of government applications, these concerns frequently involve technical difficulties, security issues, language accessibility, and usability challenges. This convergence strengthens the reliability of the findings, suggesting that issues detected across multiple models reflect core pain points for users. Thematically, these patterns align with dimensions from the SERVQUAL framework, particularly reliability, responsiveness, and assurance, indicating that these language technologies can uncover structured service quality gaps that directly impact user satisfaction and trust.

RQ1.1 How do LLMs compare to LDA in terms of coherence scores when extracting specific issues from user reviews? To what extent do the sets of identified issues overlap between the two methods?

LLMs not only perform competitively—often better—than LDA in terms of thematic coherence, but they also produce issues that map, at least in part, to LDA's discovered topics. However, LLMs appear to handle contextual nuances more flexibly, occasionally merging sub-issues that LDA treats as distinct. For practitioners, this implies that LLMs may offer more human-like representations of user concerns, supporting both broad issue discovery and the identification of deeper, interconnected themes within app reviews.

Following issue extraction, the LLMs' ability to perform multi-label classification was evaluated based on the previously identified issues. Classification confidence was quantified using Shannon entropy, which measures how decisively models assign labels to user feedback. Lower entropy values indicated confident predictions, while higher values reflected uncertainty. Entropy values above 1.5 were found to signify notable classification ambiguity. It was observed that both LDA and LLMs struggled with short reviews (e.g., "Ok"), which often lacked sufficient context for accurate issue classification, leading to higher entropy.

Both NLP techniques demonstrated similar entropy patterns overall, with an important exception: LLMs exhibited greater classification confidence when processing larger datasets, as evidenced in the DigiD application analysis. This finding suggests LLMs may offer scaling advantages for organizations with substantial feedback volumes.

RQ1.2 How do the classification confidence scores differ between LLMs and LDA when categorizing app reviews?

Both LDA and LLM-based models demonstrate similar entropy patterns, though LLMs generally produce more confident classifications across large app review datasets. Despite this, neither model consistently achieves high certainty, each shows application-specific limitations, with some contexts resulting in less reliable classifications.

To evaluate the consistency of LLM classifications across models, three agreement metrics were employed: Jensen-Shannon Divergence, Cohen's Kappa, and Krippendorff's Alpha. This multi-metric approach revealed a nuanced pattern of inter-model reliability that can have significant implications for implementation strategies.

The analysis demonstrated a methodological divergence between probabilistic and categorical agreement measures. While 1-JS Divergence scores indicated moderate alignment in the models' probabilistic issue distributions, the more rigorous categorical metrics (Cohen's Kappa and Krippendorff's Alpha) exposed more substantial classification inconsistencies. This pattern suggests that while LLMs share general thematic understanding of user concerns, they often differ significantly when required to make definitive multi-label assignments.

RQ1.3 To what extent do different LLMs agree on the categorization of issues within user reviews as measured by agreement metrics such as Krippendorf's alpha?

The level of agreement between LLMs remains moderate, especially in multi-label settings. While JS Divergence shows shared thematic understanding, label-specific consistency is limited. Interval Krippendorff's alpha offers a more nuanced view than Cohen's Kappa, revealing that even when models agree on labels, their consistency often differs.

Based on the comprehensive analysis comparing LLMs with traditional topic modeling approaches like LDA, it can be concluded that LLMs transform feedback analysis in six fundamental ways:

1. **Enhanced thematic coherence:** LLMs consistently produce more semantically coherent issue clusters than LDA, creating more meaningful and actionable categorizations that better represent user concerns.
2. **Higher classification confidence:** While neither of the approaches show a very confident classification, no matter the application, LLMs generally demonstrate more decisive classifications (particularly with larger datasets) enabling more confident decision-making.
3. **Contextual flexibility:** LLMs excel at consolidating related problems into intuitive categories while maintaining sensitivity to context, though this occasionally results in broader groupings that may obscure niche concerns that LDA might detect.
4. **Multi-model consensus:** Despite moderate inter-model agreement, LLMs collectively converge on key issues, providing reliable indicators of significant user pain points.
5. **Human-aligned interpretation:** LLMs generate issue representations that more closely resemble human interpretation, requiring less manual effort to translate findings into actions.
6. **Streamlined data processing:** Unlike LDA, which requires extensive preprocessing (tokenization, stopword removal, lemmatization), LLMs process raw feedback text directly with minimal preparation, eliminating time-consuming preprocessing steps while maintaining the contextual integrity of the original feedback.

These strengths show how advanced NLP systems address both functional and experiential dimensions of user feedback. From a theoretical perspective, several of these capabilities align closely with the SERVQUAL framework. For example, improved coherence, classification confidence, and cross-model consensus reflect greater reliability and assurance, while contextual flexibility and human-aligned interpretation relate to empathy and responsiveness in recognizing individual user concerns. Moreover, considering Uses and Gratifications Theory, these NLP-driven improvements support key user needs such as information-seeking (surveillance) or task efficiency.

RQ1

How do LLMs extract and classify issues from user generated content compared to traditional methods?

LLMs significantly enhance the extraction and classification of issues from user feedback compared to traditional methods like LDA. While categorical agreement can still pose challenges, these language technologies produce more coherent and context-sensitive topics, provide higher-confidence classifications, and offer nuanced representations that align more closely with human interpretation of user concerns.

6.1.2 RQ2

A second theoretical contribution of this thesis is the analysis of how different types of user-reported issues affect satisfaction in public service applications. RQ2 focused on evaluating the relationship between extracted issues and user ratings, as well as how the citizen concerns vary across application types and over time. This aligns with Expectation-Confirmation Theory, which views satisfaction as the result of matching user expectations with actual performance. When those expectations are unmet, it often leads to dissatisfaction, reflected in negative feedback and lower ratings. It also ties into the Technology Acceptance Model, which emphasizes that perceived usefulness and ease of use are key factors to user adoption and satisfaction.

The effect analysis reveals significant relationships between specific issue categories and user satisfaction metrics, with notable variations across different application contexts. CLMs were used to identify both universal and application-specific patterns that inform the development of targeted improvement strategies.

The CLM analysis consistently identified technical reliability and core functionality as the primary drivers of user dissatisfaction across all applications. Issues related to "Language Support" and "App Functionality Problems" demonstrated the strongest negative coefficients, indicating their substantial impact on star ratings. Conversely, some functionality-related mentions occasionally contributed to higher ratings, suggesting that constructive user feedback can positively influence satisfaction metrics even amid other concerns.

RQ2.1 What relationships can be identified between specific issue types and user satisfaction metrics, and how do these relationships vary in significance across different application contexts?

The CLM analysis highlighted both the negative and occasionally positive effects of extracted issues on review star ratings. Across all four applications, these effects vary according to each app's context. Nevertheless, technical reliability and core functionality consistently exert the strongest negative impact on satisfaction, while user interface concerns typically have more moderate negative influences that differ across applications. **KopieID** users respond most negatively to document scanning and image quality issues, **Reisapp** satisfaction is most affected by language support and stability limitations, **MijnOverheid** users primarily penalize system reliability concerns, and **DigiD** shows the strongest negative coefficients for authentication processes and stability issues. The varying significance of these issues directly reflects each application's core purpose, with users penalizing failures in primary functionality more severely than secondary features.

The time-based analysis of issue-satisfaction alignment demonstrates that user concerns shift dynamically over different periods—such as the heightened dissatisfaction during the SARS-CoV-2 pandemic—while other issues recede or become overshadowed by emerging ones. This suggests that user satisfaction is not static, but influenced by whether services meet the expectations relevant to a particular moment. These findings reflect the core of Expectation-Confirmation Theory, and extend it by showing that in digital government services, expectations are not only shaped by past use but are also highly responsive to major societal events. This dynamic also aligns with the Technology Acceptance Model, particularly in how perceptions of usefulness and ease of use influence ongoing satisfaction and adoption. As applications evolve (through updates or in response to crises) users continuously reassess these perceptions. Depending on application and timing, dissatisfaction may result from a perceived drop in usefulness (e.g., incorrect or outdated travel information), or usability (e.g., failing login systems).

RQ2.2 How can temporal analysis reveal the evolving impact of different issues on user satisfaction over time?

Temporal analysis reveals how the impact of different issues on user satisfaction evolves by combining shifts in issue frequency with changes in associated star ratings over time. For instance, the studied applications show a clear dip in satisfaction from 2020 to 2023, alongside growing appreciation for well-designed interfaces. This approach highlights that user concerns are dynamic, reinforcing the importance of ongoing monitoring to anticipate changes and guide timely improvements.

The detailed and two-front analysis of the relationship between extracted issues and user satisfaction metrics reveals multiple dimensions of influence that can directly inform continuous improvement strategies.

1. **Issue-satisfaction relationship patterns:** Technical reliability issues consistently have the strongest negative impact on satisfaction, but the specific high-impact issues vary significantly

by application purpose, with users evaluating satisfaction primarily based on an application's core functionality.

2. **Temporal evolution of issue importance:** Issue importance is not static and user priorities evolve over time, with certain issues intensifying in impact during specific periods while others diminish in importance. For government applications, technical reliability issues were observed to have consistently strong negative impacts over time, whereas usability and interface concerns showed greater variability.
3. **Supporting continuous improvement:** This multidimensional understanding of issue-satisfaction relationships supports continuous improvement by enabling more strategic resource allocation, anticipatory development prioritization, and context-sensitive enhancement planning.

RQ2

How do the extracted issues influence user satisfaction?

The influence of extracted issues on user satisfaction was quantified using coefficient estimates from CLM, combined with a temporal analysis showing how the extracted and classified issues evolve over time. The results show that issues related to technical reliability and core functionality consistently exert the strongest negative impact on star ratings. However, high-impact issues vary by app and time period: e.g., DigiD faced declining satisfaction due to authentication failures between 2020 and 2022, while KopieID was impacted by scanning issues in earlier years.

6.1.3 RQ3

The third theoretical contribution is the exploration of how language models can anticipate emerging concerns and generate actionable recommendations. Thus, RQ3 focused on assessing the predictive value of extracted issues and the usefulness of model-generated suggestions for service improvement.

A prototyped predictive framework can be refined and brought closer to completion by extending the time-analysis presented earlier with an LLM-based forecasting procedure. While historical trend analysis excels at highlighting long-standing patterns in user feedback, LLM forecasting provides a forward-looking perspective that captures newly emerging issues with greater specificity.

The forecasts demonstrate remarkable precision in identifying specific, narrowly-defined emerging issues across the government applications studied. The LLMs successfully anticipated several technical problems, including particular authentication challenges and document processing limitations, before they became widespread in user feedback. This level of granularity comes from LLMs' ability to understand nuanced context within user feedback and anticipate how identified concerns might evolve over time.

However, this focus on specific issues comes with notable limitations compared to historical trend insights. LLMs consistently overlooked broader categorical concerns that time analysis readily captures. While historical trends effectively track general issue categories like user interface problems and usability issues, LLM forecasts gravitated toward particular manifestations of these issues rather than their overarching categories, such as "UI issues with Dark Mode".

Furthermore, the LLM-based forecasts demonstrated notably limited capability in predicting the future prevalence of specific issues, aside from a few exceptions, the models struggled to gauge how widespread certain problems would become. This shortcoming reflects their emphasis on qualitative identification rather than quantitative projection: while the LLMs excel at indicating what might occur, they are less proficient at determining how often or when those issues will emerge.

The LLM approach fundamentally prioritizes issue identification over frequency prediction, making it a valuable complement to (rather than replacement for) historical trend analysis in comprehensive forecasting frameworks.

RQ3.1 How do LLM-based forecasts compare with insights derived from historical issue trends?

LLM-based forecasts offer a valuable forward-looking perspective that can uncover newly emerging issues with sharper focus, but they can underrepresent more general, recurring concerns and struggle with precise frequency estimates. Consequently, historical issue trends remain crucial for identifying long-standing patterns, while LLM forecasts add specificity and depth by spotlighting potentially overlooked or nascent problems. This complementary approach—blending past trends with future-oriented predictions—enables a more robust and proactive strategy for application improvement.

Beyond simply predicting potential problems and leaving analysts to devise solutions, these language technologies also demonstrated considerable practicality by offering targeted, app-specific recommendations. A complete set of these suggestions and insights is presented in Section 5, guiding teams to refine functionality and prevent user dissatisfaction effectively.

RQ3.2 How can LLMs be leveraged to generate product improvement suggestions for businesses?

LLM-generated forecasts provide businesses with a systematic framework to anticipate emerging user challenges through analysis of contextual patterns in feedback data. These models not only identify specific technical and functional issues before they become widespread, but can also be interactively prompted to generate concrete improvement suggestions. The effectiveness of this approach was demonstrated across all four government applications studied, where LLMs generated approximately ten specific recommendations per application that addressed both immediate technical needs and emerging user expectations.

The forecasting component of this study revealed that LLMs can indeed predict emerging issues in user feedback, offering also valuable app-specific recommendations. By detecting nuanced, potentially overlooked problems—such as new device compatibility challenges or minor usability flaws—these models provide an early warning mechanism, alerting businesses of issues before they escalate. While these forecasts show certain limitations, particularly in predicting broader concerns and issue frequency, they nonetheless enable a shift from reactive problem-solving to proactive improvement planning.

Moreover, the combination of LLM forecasts with other analyses (e.g., time-based issue-satisfaction assessment) provides a more balanced, future-oriented strategy for improving user satisfaction. This approach offers visibility into long-term trends while leveraging the predictive precision of these advanced NLP techniques to address emerging concerns proactively.

RQ3

Can LLMs forecast future issues in user feedback and provide actionable insights to help businesses address emerging challenges and improve user satisfaction?

In short, **yes**—LLMs can forecast future user feedback issues and offer actionable insights that help businesses address emerging challenges and enhance user satisfaction. By pinpointing critical vulnerabilities and proposing focused improvements, LLMs enable a proactive approach that not only reduces the likelihood of widespread dissatisfaction but also gives the possibility to create more resilient digital products.

6.2 Practical Implications

While this study focuses on four Dutch government applications, the underlying methodology and analytical framework can be generalized and scaled to user reviews across a broad range of digital platforms. The approach offers actionable insights for any organization seeking to enhance service quality through structured user feedback analysis.

6.2.1 Improving Issue Detection: Confidence-Aware and Multi-Model Strategies

One of the key practical strengths of the analyzed language technologies lies in their ability to group related user concerns into specialized categories with human-interpretable labels. For government agencies and similar institutions, this translates to more efficient issue identification and classification, enabling faster response times to critical user concerns. However, these results should be interpreted with caution, as the consolidation can sometimes produce overly broad categories that obscure less frequent, yet potentially important, issues.

A related challenge is the prevalence of high-entropy classifications, which signal uncertainty in model output. Automatically accepting uncertain labels poses a substantial risk to subsequent decision-making processes. For government agencies implementing automated feedback analysis, these findings highlight the importance of implementing confidence thresholds in classification pipelines. Businesses can improve their feedback analysis systems by requesting a succinct chain of thought alongside each assigned label and by having the model explicitly state its confidence level. This supports a confidence-weighted prioritization approach, where higher-certainty classifications receive more attention. Although implementing such mechanisms was beyond this study's scope, the identification of entropy thresholds offers guidance for building more reliable and trustworthy feedback systems.

Additionally, the observed inter-model variability highlights the need for robust validation in multi-model settings. Relying on a single LLM risks inconsistent issue detection, especially in complex datasets. Based on this analysis, several strategic approaches are recommended for organizations adopting multi-model LLM deployment. Recognizing that the initial investment in multiple models represents an accepted business cost, organizations can maximize classification reliability by: (1) implementing ensemble classification systems that synthesize outputs across models to leverage their collective intelligence and (2) establishing targeted human review protocols specifically for cases where models exhibit significant classification disagreement. These practices enable organizations to maintain the efficiency benefits of automated classification while substantially improving classification consistency and trustworthiness, ultimately enhancing the quality of insights derived from user feedback.

6.2.2 Strategic Issue Prioritization Based on Impact and Temporal Trends

The varying impact of specific issues across four government applications, as identified through the CLM analysis, underscores the need for tailored improvement strategies specific to each application:

- **KopieID** users responded most negatively to technical and performance limitations, particularly document scanning functionality and image quality problems. For this document anonymization application designed to create secure copies of documents, prioritizing improvements to core document processing capabilities would yield the most significant satisfaction gains. The direct relationship between scanning reliability and the application's primary purpose makes this the clear priority for development resources.
- **Reisapp** users demonstrated heightened sensitivity to language support limitations and interface navigation issues. As an application supporting Dutch citizens during international travel, these

concerns directly impact core functionality. Expanding language options beyond Dutch and refining core functionalities (such as the travel advice) would substantially enhance satisfaction by better serving users across diverse global locations and connectivity situations or non-Dutch speakers, such as expats.

- **MijnOverheid** users exhibited broad negative responses to reliability and functionality problems affecting this critical digital mailbox for government communications. The CLM analysis revealed system stability as the primary satisfaction determinant, directly impacting citizens' ability to access official correspondence. This points to an urgent need for infrastructure enhancements to ensure consistent availability of this essential government service.
- **DigiD** users penalized account-related and stability issues most severely, with only modest positive effects from well-executed general functionality. For the Netherlands' national digital authentication system, which serves as the gateway to virtually all government services, the findings suggest prioritizing authentication reliability and security infrastructure rather than feature expansion. These improvements would address the most critical concerns for this foundational digital identity platform.

Moreover, tracking the variation of issue frequencies and star ratings over time enables the early detection of emerging user concerns. This allows government organizations to identify early warning signals, such as rising dissatisfaction with new features or increased performance sensitivity during peak usage periods. Acting on these insights helps to proactively address pain points before they escalate into broad dissatisfaction, ensuring that development efforts remain closely aligned with real-time user sentiment.

For government applications, the temporal analysis supported findings from the effect analysis, showing that technical reliability issues exerted consistently strong negative impacts across multiple time periods, while user interface concerns varied more widely in their effect. When integrated into existing monitoring systems, time-based and sentiment analyses enable several proactive strategies: (1) prioritizing issues based on observed trajectories rather than single-point severities, (2) intervening early on accelerating problems before they reach critical levels, and (3) reallocating resources from diminishing areas of concern to emerging priorities.

6.2.3 Anticipation and Temporal Prediction in Service Response Planning

Combining historical trends from time-based analysis with the forecasting capabilities of LLMs enables organizations to develop a more comprehensive strategy, one that maintains visibility into persistent user concerns while proactively addressing emerging issues. This synergy ensures that decision-makers remain informed of persistent trends while staying agile enough to respond swiftly to changing user needs. To fully integrate LLM forecasting into their broader analytics frameworks, government institutions must balance the benefits of early issue detection with a realistic understanding of prediction limitations.

In addition to supporting early detection and strategic planning, LLM forecasting also offers concrete guidance for improvement. An examination of these application-based recommendations shows how the government can transform the forecasts into tangible improvements by focusing on three key areas:

1. **Core functionality enhancement:** Prioritize improvements to essential features that directly support the application's primary purpose. For instance, KopieID's recommendations focused on image quality and scanning reliability, which are the application's central functions.
2. **Emerging user expectation alignment:** Address shifting user expectations before they become widespread demands. The recommendations for expanding language support in Reisapp and adding biometric options to multiple applications demonstrate how businesses can stay ahead of evolving user preferences.

3. **Technical reliability assurance:** Implement proactive testing and compatibility enhancements to prevent potential issues. Recommendations for DigiD emphasized strengthening technical reliability and cross-device compatibility before wider user impact.

Thus, by implementing structured processes to generate, evaluate, and act on LLM-generated forecasts, government organizations can better levarage the capabilities of advanced NLP systems. This shift enables them to move from reactive issue resolution to proactive product enhancement. As a result, they can address emerging challenges before they impact user satisfaction and build more resilient applications that evolve with user needs.

7 Conclusion

7.1 Synthesis of Research Outcomes

This research investigated how langauage technologies, such as LLMs, can transform the analysis of user feedback to improve Dutch government applications. Four distinct applications were examined: KopieID (document anonymization), Reisapp (international travel), MijnOverheid (official digital mailbox), and DigiD (authentication), and provided answers to three main research questions through four analytical tasks: issue extraction, review classification, issue-star rating assessment, and forecasting.

The comparative analysis of various LLMs against traditional LDA demonstrated that LLMs significantly enhance feedback analysis through six key advantages: improved thematic coherence with more semantically aligned issues, higher classification confidence (despite common challenges with brief reviews), greater contextual flexibility that consolidates related problems into intuitive categories, stronger multi-model consensus as models converge on critical concerns despite individual label variations, more human-aligned interpretations, and streamlined processing that eliminates the need for extensive preprocessing. These capabilities reflect core principles from behavioral service quality theories, such as SERVQUAL, which emphasize reliability, responsiveness, and assurance in user experience evaluation.

The analysis revealed that government applications are consistently affected by recurring technical challenges, such as authentication failures, scanning errors, and broader issues related to security, language, and usability. The impact and temporal evolution of these issues on user satisfaction (measured via star ratings) were also assessed. By linking issue types to user satisfaction the study draws on Expectation-Confirmation Theory, showing that dissatisfaction arises when key expectations are unfulfilled. This dual-layered assessment underscored that issue importance is dynamic, with user priorities evolving over time. Importantly, it was demonstrated that this understanding supports continuous improvement strategies across three stages: identifying patterns in issue–satisfaction relationships, tracking temporal changes in issue importance, and enabling proactive action through resource allocation and anticipatory development prioritization.

Finally, it was shown that while historical issue analysis remains essential, it can be effectively complemented—though not fully replaced—by LLM-based forecasting, which provides a forward-looking view to identify emerging issues with greater specificity. These forecasts and recommendations can help government institutions shift reactive responses to strategic planning, focusing on functionality, evolving user expectations, and technical reliability.

Together, these key conclusions offer a valuable foundation for monitoring and enhancing user satisfaction and supporting continuous improvement of public digital services. This research demonstrates how advances in NLP, exmemplified by the use of LLMs, can shift government app feedback analysis from a reactive task to a proactive discipline that anticipates user needs and drives citizen-aligned service improvements.

The complete implementation supporting this research is publicly available and can be accessed via the following GitHub repository link: `https://github.com/Anca-Mt/THESIS`.

7.2 Limitations

This research thesis has some limitations, which are described below.

In the multi-label classification task, all evaluated LLMs exhibited hallucination tendencies, a well-documented challenge with generative AI models. In the current context, hallucinations manifested as models introducing novel issues when explicitly instructed to classify according to previously extracted issues only. This behavior may occur due to limitations in how the models interpret input, suboptimal parameter configurations, or the use of a zero-shot learning approach, where no examples are provided. Exploring alternative prompt designs, such as stricter output constraints or incorporating few-shot examples, could be a promising direction for mitigating these hallucinations.

A related constraint involves the computational cost and cost-effectiveness of LLMs, which vary by model, dataset size, and task complexity. These financial considerations required strategic decisions, including limiting Claude-3.5-Sonnet to the initial extraction task rather than extending it to classification tasks. Similarly, the forecasting evaluation was restricted to a single model to manage costs. These constraints highlight the practical challenges of deploying state-of-the-art LLMs in resource-limited research or production environments, particularly for extensive text processing pipelines spanning multiple government applications.

This study relied on evaluating model outputs generated through unsupervised methods, introducing specific methodological challenges. While LLMs offer the advantage of minimal preprocessing requirements compared to traditional NLP approaches, their evaluation required removing duplicated words from extracted issue-relevant words to prevent artificial inflation of coherence scores. Furthermore, the lack of a gold-standard labeled dataset for Dutch government applications limited the ability to assess absolute performance, making this comparative analysis between models and against LDA the primary evaluation framework.

Practical code dependency conflicts were also encountered in the pipeline. Specifically, the `googletrans` package used to translate non-Dutch reviews into Dutch required an older version of `httpx (0.19.0)` package. However, necessary AI libraries such as `mistralai` or `openai` require the newer `httpx (0.28.1)`. This incompatibility presents a significant barrier to production deployment. While this study employed workarounds suitable for research purposes, a production environment would benefit from either adopting alternative translation solutions like `deep-translate` or processing reviews in their original languages where possible. It's worth noting that while modern LLMs handle multilingual inputs effectively, the LDA baseline has inherent limitations with cross-lingual content, potentially affecting comparison validity.

A notable methodological constraint was the decision to evaluate LLMs using consistent, agnostic prompts rather than optimized ones. While this approach ensured fair comparison across models and with LDA, it likely underrepresents the full capabilities of LLMs, which typically improve with prompt engineering and refinement. The deliberate exclusion of prompt optimization techniques, such as few-shot learning, chain-of-thought prompting, or retrieval-augmented generation, from this experimental scope may have limited the observable performance. Future work incorporating systematic prompt optimization methodologies could reveal greater performance differentials between LLMs and traditional approaches like LDA.

7.3 Future Work

This research presents several promising directions for future investigation to extend and enhance the findings of this study.

Future work should explore prompt optimization techniques to improve LLM performance across issue extraction, classification, and forecasting tasks. Specifically, tailoring prompts to focus on technical issues alone, rather than general categories, could sharpen the relevance of the outputs. This specialization is expected to reduce the prevalence of broad, less actionable categories like "Usability and User Experience" or "Other", which tend to aggregate diverse concerns. This targeted approach might reveal more granular patterns in user feedback that remain obscured in general analysis.

While this study focused primarily on proprietary LLMs, expanding the experimentation to include open-source models, such as Llama-3.3-70B-Instruct, would provide valuable comparative insights. Such an extension would allow for an analysis of performance differences between proprietary and open-source models, a dimension deliberately left out of scope in the present work. Additionally, this extension would address important questions about accessibility and democratization of these analytical capabilities for government organization with varying resource constraints, and for which privacy and security are paramount.

Beyond traditional LLMs, incorporating transformer-based topic modeling approaches, such as BERTopic, would expand the evaluation framework through a broader use of modern language technologies. BERTopic has demonstrated competitive performance in extracting coherent topics across various benchmarks involving traditional models such as LDA [20], potentially offering a middle ground between traditional topic modeling and full LLM implementations.

Perhaps the most promising direction involves developing integrated human-AI collaborative systems for feedback analysis. While the current research deliberately excluded human evaluation to assess automated capabilities, creating feedback-improvement loops where human analysts guide and refine LLM outputs could significantly enhance performance. These collaborative systems could leverage human expertise for validating outputs while maintaining the efficiency advantages of automated analysis. Research on optimal interaction points for human intervention could identify where human judgment adds the most value in the feedback analysis pipeline—whether in validating issue classifications, prioritizing forecasted issues, or refining recommended improvements.

Finally, extending this research beyond public app reviews by correlating findings with application version histories and incorporating diverse feedback channels such as support tickets, user surveys, and service interactions could significantly enhance the comprehensiveness of the analysis. While government partnerships providing access to such sensitive institutional data were not established during this research, the methodology presented in this thesis underscores the potentil of NLP, with LLMs as a core component, to analyze unstructured feedback at scale. This framework could serve as a prototype for government agencies to analyze their complete feedback ecosystem, potentially revealing more nuanced patterns in citizen experiences than public reviews alone capture. Such an integrated analysis would enable more robust issue identification by combining insights from various communication channels, providing a more complete and nuanced understanding of user concerns.

References

[1] Lay Acheadeth, Misa Xirinda, and Nunung Nurul Qomariyah. "Classifying Review into Category via Topic Modelling". In: Jan. 2022.

[2] Natalia Amat-Lefort and Stuart J. Barnes. "An Inconvenient Truth: Understanding Service Inconvenience in Digital Platforms". In: *Journal of Service Research* 0.0 (0), p. 10946705241254735. DOI: 10.1177/10946705241254735. eprint: https://doi.org/10.1177/10946705241254735. URL: https://doi.org/10.1177/10946705241254735.

[3] Roy Baumeister et al. "Bad Is Stronger than Good". In: *Review of General Psychology* 5 (Dec. 2001). DOI: 10.1037/1089-2680.5.4.323.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation". In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 993–1022. ISSN: 1532-4435.

[5] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2022. arXiv: 2108.07258 [cs.LG]. URL: https://arxiv.org/abs/2108.07258.

[6] Gerlof Bouma. "Normalized (Pointwise) Mutual Information in Collocation Extraction". In: *Proceedings of the Biennial GSCL Conference 2009* (Jan. 2009).

[7] Jop Briët and Peter Harremoës. "Properties of Classical and Quantum Jensen-Shannon Divergence". In: *Physical Review A* 79 (June 2008). DOI: 10.1103/PhysRevA.79.052311.

[8] Erik Cambria et al. "Sentiment Analysis Is a Big Suitcase". In: *IEEE Intelligent Systems* 32.6 (2017), pp. 74–80. DOI: 10.1109/MIS.2017.4531228.

[9] Neil Charness and Walter R. Boot. "Chapter 20 - Technology, Gaming, and Social Networking". In: *Handbook of the Psychology of Aging (Eighth Edition)*. Ed. by K. Warner Schaie and Sherry L. Willis. Eighth Edition. San Diego: Academic Press, 2016, pp. 389–407. ISBN: 978-0-12-411469-2. DOI: https://doi.org/10.1016/B978-0-12-411469-2.00020-0. URL: https://www.sciencedirect.com/science/article/pii/B9780124114692000200.

[10] *ChatGPT's Architecture*. 2024. URL: https://www.geeksforgeeks.org/chatgpts-architecture/.

[11] Ning Chen et al. "AR-miner: mining informative reviews for developers from mobile app marketplace". In: *Proceedings of the 36th International Conference on Software Engineering*. ICSE 2014. Hyderabad, India: Association for Computing Machinery, 2014, pp. 767–778. ISBN: 9781450327565. DOI: 10.1145/2568225.2568263. URL: https://doi.org/10.1145/2568225.2568263.

[12] Wei-Lin Chiang et al. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. 2024. arXiv: 2403.04132 [cs.AI]. URL: https://arxiv.org/abs/2403.04132.

[13] Adelina Ciurumelea et al. "Analyzing reviews and code of mobile apps for better release planning". In: *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. 2017, pp. 91–102. DOI: 10.1109/SANER.2017.7884612.

[14] Mamata Das, Selvakumar K., and P. J. A. Alphonse. *A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset*. 2023. arXiv: 2308.04037 [cs.CL]. URL: https://arxiv.org/abs/2308.04037.

[15] Fred D. Davis. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology". In: *MIS Quarterly* 13.3 (1989), pp. 319–340. ISSN: 02767783, 21629730. URL: http://www.jstor.org/stable/249008 (visited on 05/10/2025).

[16] Taha Falatouri, Denisa Hrušecká, and Thomas Fischer. “Harnessing the Power of LLMs for Service Quality Assessment From User-Generated Content”. In: *IEEE Access* PP (Jan. 2024), pp. 1–1. DOI: 10.1109/ACCESS.2024.3429290.

[17] GENSIM. *Topic modelling for humans*. 2024. URL: https://radimrehurek.com/gensim/.

[18] P. Greeshma, Jisha Vijay, and P. Mohan Kumar. “THE EXPECTATION CONFIRMATION THEORY: A SERVICE PERSPECTIVE”. In: *GAP Bodhi Taru: A Global Journal of Humanities* 8 (Mar. 2025). URL: https://www.gapbodhitaru.org/res/articles/(10-18)%20THE%20EXPECTATION%20CONFIRMATION%20THEORY%20A%20SERVICE%20PERSPECTIVE.pdf.

[19] Andreas Gregoriades et al. “Explaining tourist revisit intention using natural language processing and classification techniques”. In: *Journal of Big Data* 10 (May 2023). DOI: 10.1186/s40537-023-00740-5.

[20] Maarten Grootendorst. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. 2022. arXiv: 2203.05794 [cs.CL]. URL: https://arxiv.org/abs/2203.05794.

[21] Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. *Comprehensive Study on Sentiment Analysis: From Rule-based to modern LLM based system*. 2024. arXiv: 2409.09989 [cs.CL]. URL: https://arxiv.org/abs/2409.09989.

[22] Valentinus Roby Hananto, Uwe Serdült, and Victor Kryssanov. “A Text Segmentation Approach for Automated Annotation of Online Customer Reviews, Based on Topic Modeling”. In: *Applied Sciences* 12.7 (2022). ISSN: 2076-3417. DOI: 10.3390/app12073412. URL: https://www.mdpi.com/2076-3417/12/7/3412.

[23] Rune Haubo and Bojesen Christensen. “Cumulative Link Models for Ordinal Regression with the R Package ordinal”. In: 2018. URL: https://api.semanticscholar.org/CorpusID:59572956.

[24] Thomas Horan, Tarun Abhichandani, and R. Rayalu. “Assessing User Satisfaction of E-Government Services: Development and Testing of Quality-in-Use Satisfaction with Advanced Traveler Information Systems (ATIS)”. In: vol. 4. Feb. 2006, 83b–83b. ISBN: 0-7695-2507-5. DOI: 10.1109/HICSS.2006.66.

[25] IBM. *What are large language models (LLMs)?* 2023. URL: https://www.ibm.com/think/topics/large-language-models.

[26] Medium: Abhishek Jain. *TF-IDF in NLP (Term Frequency Inverse Document Frequency)*. 2024. URL: https://medium.com/%40abhishekjainindore24/tf-idf-in-nlp-term-frequency-inverse-document-frequency-e05b65932f1d.

[27] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released January 12, 2025. 2025. URL: https://web.stanford.edu/~jurafsky/slp3/.

[28] Elihu Katz, Jay G. Blumler, and Michael Gurevitch. “Uses and Gratifications Research”. In: *The Public Opinion Quarterly* 37.4 (1973), pp. 509–523. ISSN: 0033362X, 15375331. URL: http://www.jstor.org/stable/2747854 (visited on 05/10/2025).

[29] Klaus Krippendorff. *Computing Krippendorff’s Alpha-Reliability*. 2011. URL: https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf.

[30] Kleopatra Lapis and Allah Ditta. *Benchmarking LLMs in E-commerce: Sentiment Analysis and Causal Reasoning in Customer Feedback*. Aug. 2024. DOI: 10.13140/RG.2.2.24875.25120.

[31] Yi Liu et al. “Sentiment analysis for e-commerce product reviews by deep learning model of Bert-BiGRU-Softmax”. In: *Mathematical Biosciences and Engineering* 17 (Nov. 2020), pp. 7819–7837. DOI: 10.3934/mbe.2020398.

[32] Mary McHugh. “Interrater reliability: The kappa statistic”. In: *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB* 22 (Oct. 2012), pp. 276–82. DOI: 10.11613/BM.2012.031.

[33] Meta. *Introducing Meta Llama 3: The most capable openly available LLM to date*. 2024. URL: https://ai.meta.com/blog/meta-llama-3/.

[34] Christian Morbidoni. “Poster: LLMs for online customer reviews analysis: oracles or tools? Experiments with GPT 3.5”. In: Sept. 2023, pp. 1–4. DOI: 10.1145/3605390.3610810.

[35] mouseflow. *User Feedback: The Ultimate Guide*. 2025. URL: https://mouseflow.com/topics/user-feedback/.

[36] Mekhail Mustak et al. “Using machine learning to develop customer insights from user-generated content”. In: *Journal of Retailing and Consumer Services* 81 (2024), p. 104034. ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2024.104034. URL: https://www.sciencedirect.com/science/article/pii/S0969698924003308.

[37] Government of the Netherlands. *DigiD*. 2017. URL: https://www.digid.nl/.

[38] Government of the Netherlands. *MijnOverheid*. 2018. URL: https://mijn.overheid.nl/.

[39] Government of the Netherlands - National Office for Identity Data. *KopieID*. 2014. URL: https://www.rvig.nl/kopieid-app.

[40] Government of the Netherlands - Rijksoverheid. *Reisapp*. 2012. URL: https://www.rijksappstore.nl/reisapp.

[41] Thuy Ngoc Nguyen, Kasturi Jamale, and Cleotilde Gonzalez. *Predicting and Understanding Human Action Decisions: Insights from Large Language Models and Cognitive Instance-Based Learning*. 2024. arXiv: 2407.09281 [cs.AI]. URL: https://arxiv.org/abs/2407.09281.

[42] Richard Oliver. “Effect of Expectation and Disconfirmation on Postexposure Product Evaluations: An Alternative Interpretation”. In: *Journal of Applied Psychology* 62 (Aug. 1977), pp. 480–486. DOI: 10.1037/0021-9010.62.4.480.

[43] Bo Pang and Lillian Lee. 2008. DOI: 10.1561/1500000011.

[44] A Parsu Parasuraman, Valarie Zeithaml, and Leonard Berry. “SERVQUAL A Multiple-item Scale for Measuring Consumer Perceptions of Service Quality”. In: *Journal of Retailing* 64 (Jan. 1988), pp. 12–40.

[45] S.V. Praveen et al. “Crafting clarity: Leveraging large language models to decode consumer reviews”. In: *Journal of Retailing and Consumer Services* 81 (2024), p. 103975. ISSN: 0969-6989. DOI: https://doi.org/10.1016/j.jretconser.2024.103975. URL: https://www.sciencedirect.com/science/article/pii/S0969698924002716.

[46] Juan Ramos. “Using TF-IDF to determine word relevance in document queries”. In: (Jan. 2003).

[47] Abu Rayhan, Robert Kinzler, and Rajan Rayhan. *NATURAL LANGUAGE PROCESSING: TRANSFORMING HOW MACHINES UNDERSTAND HUMAN LANGUAGE*. Aug. 2023. DOI: 10.13140/RG.2.2.34900.99200.

[48] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: http://arxiv.org/abs/1908.10084.

[49] Michael Röder, Andreas Both, and Alexander Hinneburg. "Exploring the Space of Topic Coherence Measures". In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15. Shanghai, China: Association for Computing Machinery, 2015, pp. 399–408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324. URL: https://doi.org/10.1145/2684822.2685324.

[50] SBERT.net. *Sentence Transformer*. 2019. URL: https://sbert.net/docs/quickstart.html.

[51] Towards Data Science. *Cosine similarity: How does it measure the similarity, Maths behind and usage in Python*. 2020. URL: https://towardsdatascience.com/cosine-similarity-how-does-it-measure-the-similarity-maths-behind-and-usage-in-python-50ad30aad7db/.

[52] Murray Scott, William Golden, and Martin Hughes. "The Implementation Of Citizen-Centred E-government: a Stakeholder Viewpoint". In: (Mar. 2004).

[53] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[54] Minghao Shao et al. "Survey of Different Large Language Model Architectures: Trends, Benchmarks, and Challenges". In: *IEEE Access* 12 (2024), pp. 188664–188706. ISSN: 2169-3536. DOI: 10.1109/access.2024.3482107. URL: http://dx.doi.org/10.1109/ACCESS.2024.3482107.

[55] Xiaoming Shi et al. *Language Models Can Improve Event Prediction by Few-Shot Abductive Reasoning*. 2023. arXiv: 2305.16646 [cs.CL]. URL: https://arxiv.org/abs/2305.16646.

[56] UC Berkeley SkyLab and LMArena. *Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots*. 2025. URL: https://lmarena.ai/?leaderboard.

[57] Jing Su et al. *Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review*. 2024. arXiv: 2402.10350 [cs.LG]. URL: https://arxiv.org/abs/2402.10350.

[58] Shaheen Syed and Marco Spruit. "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation". In: Oct. 2017, pp. 165–174. DOI: 10.1109/DSAA.2017.61.

[59] Hua Tang et al. *Time Series Forecasting with LLMs: Understanding and Enhancing Model Capabilities*. 2024. arXiv: 2402.10835 [cs.CL]. URL: https://arxiv.org/abs/2402.10835.

[60] DATAtab Team. *Cohen's Kappa*. 2025. URL: https://datatab.net/tutorial/cohens-kappa.

[61] TiDB. *Understanding the Cosine Similarity Formula*. 2024. URL: https://www.pingcap.com/article/understanding-the-cosine-similarity-formula/.

[62] Grigorios Tsoumakas and Ioannis Katakis. "Multi-Label Classification: An Overview". In: *International Journal of Data Warehousing and Mining* 3 (Sept. 2009), pp. 1–13. DOI: 10.4018/jdwm.2007070101.

[63] Department of Economic United Nations and Social Affairs. *UN E-Government Survey 2022*. 2022. URL: https://publicadministration.un.org/egovkb/en-us/Reports/UN-E-Government-Survey-2022.

[64] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[65] Zhiqiang Wang, Yiran Pang, and Yanbin Lin. *Smart Expert System: Large Language Models as Text Classifiers*. May 2024. DOI: 10.48550/arXiv.2405.10523.

[66] Jialiang Wei et al. "Zero-shot Bilingual App Reviews Mining with Large Language Models". In: *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, Nov. 2023, pp. 898–904. DOI: 10.1109/ictai59109.2023.00135. URL: http://dx.doi.org/10.1109/ICTAI59109.2023.00135.

[67] Nour Eddine Zekaouiu et al. "Analysis of the evolution of advanced transformer-based language models: experiments on opinion mining". In: *IAES International Journal of Artificial Intelligence (IJ-AI)* 12.4 (Dec. 2023), p. 1995. ISSN: 2089-4872. DOI: 10.11591/ijai.v12.i4.pp1995-2010. URL: http://dx.doi.org/10.11591/ijai.v12.i4.pp1995-2010.

[68] Wayne Xin Zhao et al. *A Survey of Large Language Models*. 2025. arXiv: 2303.18223 [cs.CL]. URL: https://arxiv.org/abs/2303.18223.

A Appendix CLM Statistics

Tables 7 - 10 present the estimated effects of individual issues on the app's star ratings, based on an ordinal regression model. The results marked with an asterisk (*) are considered statistically significant, meaning their p-value is not greater than 0.05. Results without this mark are considered not significant.

Issue	*Coef.*	*Std. Err.*	*z*	$P > \|z\|$	*[0.025*	*0.975*]
App Functionality and Performance Issues	-1.5772	0.228	-6.909	0.000*	-2.025	-1.130
Image Quality Issues	-0.8818	0.227	-3.877	0.000*	-1.328	-0.436
User Experience and Usability	0.1593	0.210	0.759	0.448	-0.252	0.571
Scanning Issues	-0.8602	0.253	-3.400	0.001*	-1.356	-0.364
Incorrect Masking/Redaction	-0.1987	0.188	-1.055	0.291	-0.568	0.170
Saving and Exporting Issues	0.0362	0.192	0.189	0.850	-0.339	0.412
Update Related Problems	-0.3390	0.230	-1.473	0.141	-0.790	0.112
Watermark Issues	-0.2219	0.301	-0.737	0.461	-0.812	0.368
Identification Card and Passport Compatibility	-0.2055	0.226	-0.909	0.363	-0.649	0.238

Note: * $p < 0.05$

Table 7: CLM Regression Statistics Output for KopieID

Issue	*Coef.*	*Std. Err.*	*z*	$P > \|z\|$	*[0.025*	*0.975*]
App Stability and Performance	-2.4951	0.367	-6.803	0.000*	-3.214	-1.776
User Interface and Usability	0.6496	0.333	1.950	0.051	-0.003	1.302
Information Accuracy and Updates	0.1934	0.320	0.604	0.546	-0.434	0.821
Missing Features/Functionality	-1.3972	0.396	-3.531	0.000*	-2.173	-0.622
Irrelevant/Unnecessary Information	-1.7355	0.504	-3.442	0.001*	-2.724	-0.747
Language Support	-2.8321	0.383	-7.396	0.000*	-3.583	-2.082
Incorrect Travel Advice	-2.2626	0.601	-3.767	0.000*	-3.440	-1.085
App Permissions	-4.5657	2.747	-1.662	0.097	-9.951	0.819

Note: * $p < 0.05$

Table 8: CLM Regression Statistics Output for Reisapp

Issue	*Coef.*	*Std. Err.*	*z*	$P > \|z\|$	*[0.025*	*0.975*]
DigiD Login Issues	-1.7476	0.213	-8.189	0.000*	-2.166	-1.329
App Functionality Problems	-3.3898	0.215	-15.746	0.000*	-3.812	-2.968
User Interface and Experience Issues	-0.7753	0.214	-3.619	0.000*	-1.195	-0.355
Content and Coverage Limitations	-1.9203	0.350	-5.494	0.000*	-2.605	-1.235
General Dissatisfaction	-2.6915	0.218	-12.358	0.000*	-3.118	-2.265

Note: * $p < 0.05$

Table 9: CLM Regression Statistics Output for MijnOverheid

Issue	*Coef.*	*Std. Err.*	*z*	$P > \lvert z \rvert$	*[0.025*	*0.975*]
Scanning Issues	-1.1149	0.063	-17.589	0.000*	-1.239	-0.991
Document Recognition Problems	-0.3780	0.142	-2.665	0.008*	-0.656	-0.100
App Stability and Bugs	-1.5469	0.069	-22.350	0.000*	-1.683	-1.411
General Functionality	0.3413	0.073	4.696	0.000*	0.199	0.484
Usability and User Experience	0.0108	0.067	0.161	0.872	-0.120	0.142
Login Problems	-0.4060	0.064	-6.368	0.000*	-0.531	-0.281
Update Issues	-0.3294	0.085	-3.895	0.000*	-0.495	-0.164
Language Barrier	-0.3884	0.140	-2.781	0.005*	-0.662	-0.115
Privacy Concerns	-0.7099	0.153	-4.629	0.000*	-1.010	-0.409
Account Issues	-1.2244	0.112	-10.980	0.000*	-1.443	-1.006
SMS Verification Problems	-0.2340	0.134	-1.747	0.081	-0.497	0.029
Server Communication Issues	-0.5046	0.139	-3.640	0.000*	-0.776	-0.233
Location Based Restriction	-0.9682	0.178	-5.455	0.000*	-1.316	-0.620

Note: * $p < 0.05$

Table 10: CLM Regression Statistics Output for DigiD