# Unveiling Dual Quality in Product Reviews: An NLP-Based Approach

**Rafał Poświata, Marcin Michał Mirończuk, Sławomir Dadas,**
**Małgorzata Grębowiec, Michał Perełkiewicz**
National Information Processing Institute
al. Niepodległości 188b, 00-608 Warsaw, Poland
{rposwiata, mmironczuk, sdadas, mgrebowiec, mperelkiewicz}@opi.org.pl

## Abstract

Consumers often face inconsistent product quality, particularly when identical products vary between markets, a situation known as the dual quality problem. To identify and address this issue, automated techniques are needed. This paper explores how natural language processing (NLP) can aid in detecting such discrepancies and presents the full process of developing a solution. First, we describe in detail the creation of a new Polish-language dataset with 1,957 reviews, 540 highlighting dual quality issues. We then discuss experiments with various approaches like SetFit with sentence-transformers, transformer-based encoders, and LLMs, including error analysis and robustness verification. Additionally, we evaluate multilingual transfer using a subset of opinions in English, French, and German. The paper concludes with insights on deployment and practical applications.

## 1 Introduction

Dual quality of products refers to practices where companies sell items under the same brand and similar packaging in different markets, yet present them with significantly altered composition or quality parameters (The European Consumer Organisation (BEUC), 2018). This phenomenon has sparked growing controversy among consumers, especially within the European Union (EU), where it is perceived as a potential violation of fair competition rules (The European Consumer Organisation (BEUC), 2018). From a sociological and economic perspective, dual quality practices raise multifaceted concerns about market trust, purchasing behaviours and the perception of fairness among consumers (Veselovská, 2022; Bartkova and Sirotiaková, 2021). Multiple reports published by consumer organizations and EU research services suggest that offering products with distinct ingredients or characteristics under identical branding constitutes a widespread international issue (The European Consumer Organisation (BEUC), 2018; European Parliament, 2019; European Commission, 2023). The above reasons and EU regulations—such as the amended Directive on Unfair Commercial Practices—recognize dual quality as misleading conduct, which may require enforcement at the national level (Chambers; EU Monitor) (also, see more details in Appendix A). Our recent research project focused on creating a solution to support a national agency from one of the EU countries to address the above problem, namely the Office of Competition and Consumer Protection (UOKiK) in Poland (`https://uokik.gov.pl/en`).

The main goal of the project was to automate the detection of unfair commercial practices using natural language processing (NLP) methods. The project, currently in the proof-of-concept stage, is enabling the automated collection and analysis of product-related data from e-commerce sites and social media. It comprises a data retrieval module (intelligent web crawling, scraping, cleaning, and preprocessing) and a text analysis module that includes language identification, sentiment analysis, aspect base sentiment analysis, and the detection of consumer reviews[1] that may indicate potential dual quality issues in products.

In this paper, we focus on the last and most novel of these components for detecting dual quality reviews, describing the entire process from data preparation, through extensive evaluation of different approaches, to deployment. To our knowledge, no available dataset or model is aimed at recognizing dual quality-related reviews. While several articles (discussed further in Section 2) approach

[1]In this article, we use the terms 'reviews' and 'opinions' interchangeably to refer to consumer expressions regarding a product. While 'review' may often imply a structured evaluation, we also include informal opinions that may indicate perceptions of dual quality.



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1480–1497
July 28-30, 2025 ©2025 Association for Computational Linguistics

dual quality from sociological, economic, and legal perspectives, our study takes a different approach presented in Figure 1.

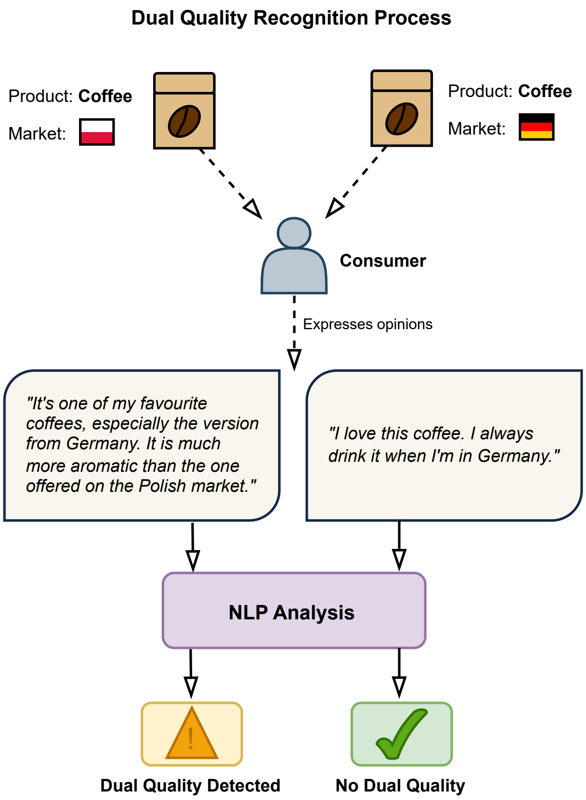


Figure 1: Illustration of the NLP-based workflow for recognizing dual quality consumer reviews. The dual quality detection system flags reviews for potential issues when a consumer explicitly notes a difference between product versions from different markets. This illustration exemplifies the process with a Polish consumer assessing products from Polish and German markets; the reviews shown are English translations of the original Polish texts for clarity and wider accessibility.

The main contributions of this work can be summarized as follows:

– Proposition of new NLP task: detecting the dual quality issues in product reviews.
– A coherent methodology for dataset construction and preparation of a corpus of 1,957 human-verified product reviews, 540 of which potentially exhibit dual quality.
– A comprehensive evaluation of Polish and multilingual models, including a presentation of various metrics, error analysis, and robustness verification conducted primarily for Polish.
– Expansion of the dataset to include product reviews in other key languages such as English, German, and French, demonstrating the system's multilingual capabilities.

## 2 Related Work

Economic and social research on dual quality products highlights the erosion of consumer trust when identical branding masks disparities in product quality across EU Member States. Studies indicate that these discrepancies, particularly in food products, impact consumer perceptions of fairness and lead to behavioral changes in purchasing decisions (Bartková et al., 2018; Bartková, 2019; Bartkova et al., 2021; Bartkova and Sirotiaková, 2021). Research has further demonstrated that wealthier consumers are more aware of the issue and seek alternatives in other markets, whereas lower-income consumers are more likely to adapt their behavior to avoid lower-quality products (Bartkova and Sirotiaková, 2021). The perception of dual quality as an economic problem is also evident, as lower-quality ingredients often correspond to price disparities that disadvantage consumers in specific regions (Závadský and Hiadlovský, 2020).

Additionally, empirical studies confirm that public perception of dual quality is shaped by exposure to media reports and political discourse, leading to heightened scrutiny of multinational corporations and their regional product differentiation strategies (Veselovská, 2022). While some scholars argue that manufacturers may justify product variations based on local market preferences, research suggests that these practices often lack transparency and leave consumers feeling deceived (Bartkova and Veselovska, 2023). Moreover, comparative consumer tests confirm that dual quality is not confined to food products but also extends to household and personal care items, reinforcing the need for regulatory intervention (Bartková and Veselovská, 2024). Given the strong consumer opposition across Europe, particularly in Central and Eastern European countries, economic research increasingly supports regulatory measures to curb these practices and ensure consistent product quality across EU markets.

From an computer science perspective, the topic of applying NLP techniques to e-commerce platforms and customer behavior analysis is widely studied. Among these works, we can point out customer reviews analysis (Botunac et al., 2024; Satjathanakul and Siriborvornratanakul, 2024; Mamani-Coaquira and Villanueva, 2024), product question answering (Shen et al., 2023; Wang et al., 2023), product categorization (Gong et al., 2023), moderation of e-commerce reviews (Nayak and Garera, 2022), product feature extraction from the web (Fuchs et al., 2022), customer service support (Obadinma et al., 2022), data augmentation in e-commerce (Avigdor et al., 2023), fake news detection (Hu et al., 2023), predictive quality in manufacturing (Tercan and Meisen, 2022), or intent classification (Parikh et al., 2023). However, none of these works address the dual quality problem directly or consider how to harness consumer opinions—such as reviews from the Internet, e-commerce platforms, or social media—to help resolve this issue. Thus, a clear research gap exists in applying NLP-based methods to detect or analyze dual quality products.

## 3 DQ Dataset

### 3.1 Dataset Creation Methodology

In the first stage of our work, we collected a large dataset of reviews in Polish, sourced from the e-commerce platform CENEO[2] and the discussion forum on beauty, makeup, and cosmetics, WIZAZ[3]. Our preliminary tests have shown that the problem of dual quality does not occur often in reviews, and thus randomly selecting a set of opinions and giving them to annotators is an inefficient approach to building a dataset. Therefore, we prepared a methodology to optimize this process, which consists of the following steps:

① Find dual quality reviews on the Internet by searching for publicly available articles that describe the problem of dual quality. Such articles often included examples of products along with the differences observed depending on the sales market, which we extracted. In addition, some articles had comment sections where people shared their experiences with the dual quality issue, which we also collected. In this way, we obtained **117** dual quality reviews.

② Randomly select **300** reviews from the CENEO / WIZAZ dataset as standard opinions that do not indicate a dual quality problem. These reviews have been verified to ensure that they are standard. Along with the examples obtained in step ①, these formed the base dataset.

③ Train a model using a few-shot learning method to detect dual quality reviews based on the prepared base or an extended dataset (subsequent iterations). We adopted this approach due to the limited amount of training data. The model was implemented using the SetFit (Sentence Transformer Fine-tuning) framework (Tunstall et al., 2022) and a sentence transformer for the Polish language st-polish-paraphrase-from-distilroberta[4].

④ Apply the model trained in step ③ to all reviews of the CENEO / WIZAZ dataset. The results of the classification were sorted according to the probability returned by the model.

⑤ Select up to **200**[5] reviews with the highest probability of indicating a dual quality problem, which did not appear previously in the dataset. Then perform manual verification of the selected reviews. If a review did not indicate a dual quality issue, it was labeled as a standard review. During this step, we noticed that some reviews mentioned other problems, including, for example, the product being possibly counterfeit, deterioration in product quality over time, or the received product does not match the order. Annotators labeled such opinions as other problems and added additional information regarding the type of problem mentioned in the review. For training the model in step ③, the reviews labeled as other problems and standard were combined. The outcome of this step and the base dataset constituted the extended dataset.

⑥ Return to step ③ to increase the size of the dataset.

Steps ③, ④, and ⑤ were repeated **7** times, allowing us to expand the base dataset with **1,303** examples (in last iteration only **103** new reviews were selected). We then applied the model, trained on the entire dataset prepared so far, to classify the reviews imported into the demo version of our system. Reviews were sourced from Polish and international e-commerce sites. Of these reviews, **237** were labeled as dual quality, which we manually verified and changed if necessary. As a result of the entire process described above, we obtained a DQ (**D**ual **Q**uality) dataset consisting of **1,957** unique examples. o ensure annotation accuracy, we conducted cross-validation and identified examples where the models were most often wrong. After verifying these errors, in **67 (3.4%)** cases the label was incorrect and was changed. The whole above process is shown in Figure 4.

[2] https://www.ceneo.pl/

[3] https://wizaz.pl/forum/

[4] At the time of the dataset creation (beginning of 2023) it was the top Polish sentence transformer, as confirmed by Dadas et al. (2024b).

[5] Initially, many reviews were classified as dual quality, making a probability threshold unsuitable. Selecting 200 enabled swift human verification, speeding up subsequent iterations.

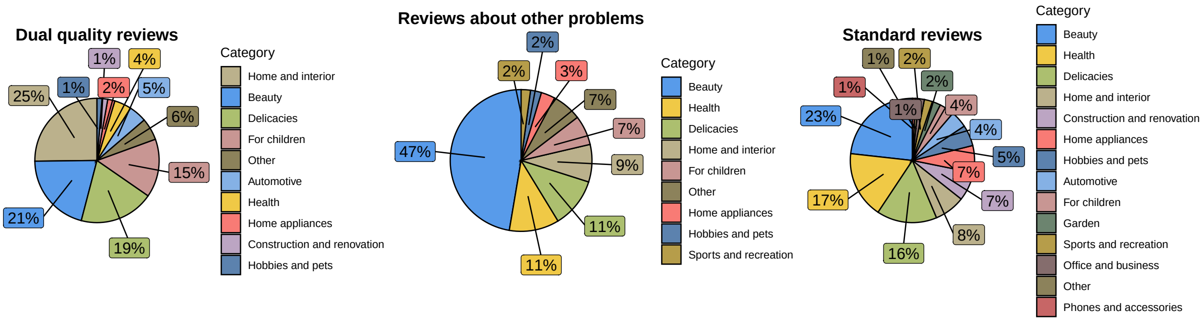


Figure 2: Charts illustrating the distribution of product categories across various types of reviews.

### 3.2 Dataset Statistics

The statistics of the DQ dataset are presented in Table 1. The dataset consists of **1,957** records, of which **540** are labeled as dual quality, **281** as other problems, and the rest are standard opinions. Of the dual quality reviews, **107**[6] were from the Internet, **265** from the CENEO / WIZAZ collection, and **168** from our demo system. The dataset is unbalanced, with over half of the reviews belong to the standard class. This characteristic was intentionally maintained because, in the real world, reviews on dual quality and other problems occur less frequently than others. For experimental purposes, the dataset was divided into three subsets: train, test and valid, containing **1,200 (∼61%)**, **500 (∼26%)**, and **257 (∼13%)** reviews, respectively. The review texts in the dataset consist of **261** characters and **41** words on average.

| | # reviews | | | |
|---|---|---|---|---|
| **label** | **all** | **train** | **test** | **valid** |
| **dual quality** | 540 | 331 | 138 | 71 |
| **other problems** | 281 | 172 | 72 | 37 |
| **standard** | 1136 | 697 | 290 | 149 |
| **total** | 1957 | 1200 | 500 | 257 |

Table 1: DQ dataset statistics.

In addition, in Figure 2 we present pie charts depicting the distribution of product categories across various types of reviews[7]. A few interesting patterns in these distributions are worth describing. For instance, although *Beauty*, *Delicacies*, *Health*, and *Home & Interior* are large categories overall, *Home & Interior* has an exceptionally high share among dual quality reviews (25%, compared to 13% overall), suggesting that this type of issue might be more commonly perceived in products related to household items. Similarly, *For children* makes up only 7% of all reviews but appears more prominently (15%) in dual quality reviews. Meanwhile, *Beauty* reviews account for nearly half (47%) of the 'other problems' category, indicating that consumers in that segment may encounter a broader range of product issues beyond dual quality concerns.

[6]In the results of the final dataset verification, of the 117 dual quality reviews initially found, 10 were classified as standard.

[7]All product reviews categorized by product type reader may see in Figure 6.

## 4 Experiments

### 4.1 Experimental Setup

The problem was defined as a three-class classification (see Table 1). Evaluation of various methods was performed on a test set. The training set and the validation set were used for approaches that required training/fine-tuning. Each experiment was repeated five times[8], setting a different seed value (if applicable), and the results presented in the tables are average values.

### 4.2 Methods

**Baseline** is a naive method of assigning a dual quality class to a review if there are references to another country in the text.

**SetFit + sentence transformers** is an approach in which a sentence transformer model is first fine-tuned using contrastive learning and then used as text embedding for a logistic regression model. In the experiments, we used sentence transformers previously tested on the PL-MTEB benchmark by Poświata et al. (2024). We selected seven multilingual models namely: LaBSE (Feng et al., 2022), paraphrase-multilingual-mpnet-base-v2, paraphrase-multilingual-MiniLM-L12-v2

[8]This rule was not applied to Baseline, which is deterministic, and successive runs always produce the same result.

| | Dual Quality class | | | All classes | | | |
|---|---|---|---|---|---|---|---|
| Method | Precision | Recall | F1 | Accuracy | mPrecision | mRecall | mF1 |
| Baseline | $42.4_{\pm 0.0}$ | $84.8_{\pm 0.0}$ | $56.5_{\pm 0.0}$ | $55.2_{\pm 0.0}$ | $37.8_{\pm 0.0}$ | $46.5_{\pm 0.0}$ | $39.5_{\pm 0.0}$ |
| **SetFit + sentence transformers** | | | | | | | |
| LaBSE | $74.4_{\pm 1.0}$ | $71.4_{\pm 2.2}$ | $72.9_{\pm 1.1}$ | $77.7_{\pm 0.5}$ | $\mathbf{75.6}_{\pm 0.8}$ | $65.9_{\pm 0.9}$ | $68.4_{\pm 0.7}$ |
| para-multi-mpnet-base-v2 | $72.8_{\pm 1.7}$ | $66.4_{\pm 2.4}$ | $69.4_{\pm 2.0}$ | $75.9_{\pm 1.4}$ | $72.4_{\pm 2.2}$ | $66.8_{\pm 2.5}$ | $68.8_{\pm 2.6}$ |
| para-multi-MiniLM-L12-v2 | $69.4_{\pm 2.2}$ | $58.7_{\pm 3.3}$ | $63.6_{\pm 2.7}$ | $71.2_{\pm 1.2}$ | $65.8_{\pm 1.3}$ | $58.2_{\pm 1.7}$ | $60.2_{\pm 1.7}$ |
| multi-e5-small | $68.7_{\pm 1.6}$ | $68.0_{\pm 1.3}$ | $68.3_{\pm 0.8}$ | $72.8_{\pm 0.7}$ | $70.4_{\pm 0.8}$ | $58.9_{\pm 0.9}$ | $60.3_{\pm 1.3}$ |
| multi-e5-base | $72.2_{\pm 1.2}$ | $\mathbf{79.0}_{\pm 2.5}$ | $75.4_{\pm 0.8}$ | $77.4_{\pm 1.0}$ | $73.7_{\pm 2.1}$ | $67.6_{\pm 1.8}$ | $69.0_{\pm 1.9}$ |
| multi-e5-large | $77.5_{\pm 1.8}$ | $76.8_{\pm 3.4}$ | $\mathbf{77.1}_{\pm 2.4}$ | $\mathbf{79.6}_{\pm 1.8}$ | $75.2_{\pm 2.8}$ | $71.2_{\pm 2.2}$ | $\mathbf{72.7}_{\pm 2.2}$ |
| gte-multi-base | $73.4_{\pm 1.1}$ | $\mathbf{79.0}_{\pm 3.4}$ | $76.1_{\pm 2.2}$ | $78.6_{\pm 0.8}$ | $74.3_{\pm 1.1}$ | $69.4_{\pm 2.0}$ | $70.8_{\pm 1.7}$ |
| st-polish-para-mpnet | $72.5_{\pm 2.0}$ | $71.7_{\pm 3.3}$ | $72.1_{\pm 2.6}$ | $76.6_{\pm 1.1}$ | $72.2_{\pm 1.3}$ | $68.1_{\pm 2.1}$ | $69.6_{\pm 1.8}$ |
| st-polish-para-distilroberta | $72.7_{\pm 2.7}$ | $69.1_{\pm 2.7}$ | $70.9_{\pm 2.6}$ | $75.7_{\pm 0.7}$ | $70.5_{\pm 0.3}$ | $68.1_{\pm 1.6}$ | $69.1_{\pm 1.1}$ |
| mmlw-roberta-base | $\mathbf{77.9}_{\pm 0.8}$ | $73.6_{\pm 1.6}$ | $75.7_{\pm 0.5}$ | $78.6_{\pm 0.6}$ | $73.4_{\pm 1.1}$ | $71.9_{\pm 1.0}$ | $72.6_{\pm 1.0}$ |
| mmlw-roberta-large | $76.0_{\pm 1.9}$ | $75.9_{\pm 2.4}$ | $75.9_{\pm 2.0}$ | $78.7_{\pm 1.4}$ | $72.7_{\pm 1.8}$ | $\mathbf{72.1}_{\pm 1.7}$ | $72.4_{\pm 1.7}$ |
| **Transformer-based encoders** | | | | | | | |
| mBERT | $64.8_{\pm 2.7}$ | $67.5_{\pm 2.0}$ | $66.1_{\pm 1.6}$ | $71.1_{\pm 1.9}$ | $62.5_{\pm 9.4}$ | $58.3_{\pm 3.5}$ | $58.6_{\pm 5.5}$ |
| xlm-roberta-base | $60.7_{\pm 1.5}$ | $82.2_{\pm 3.6}$ | $69.8_{\pm 1.1}$ | $73.1_{\pm 0.8}$ | $70.6_{\pm 1.1}$ | $63.0_{\pm 2.3}$ | $62.8_{\pm 2.5}$ |
| xlm-roberta-large | $78.3_{\pm 3.0}$ | $86.1_{\pm 2.0}$ | $\mathbf{82.0}_{\pm 1.5}$ | $82.0_{\pm 1.2}$ | $75.8_{\pm 1.7}$ | $\mathbf{76.4}_{\pm 1.6}$ | $75.9_{\pm 1.6}$ |
| herbert-base-cased | $64.0_{\pm 3.9}$ | $77.8_{\pm 3.3}$ | $70.1_{\pm 1.6}$ | $73.3_{\pm 0.2}$ | $77.3_{\pm 3.3}$ | $59.9_{\pm 2.3}$ | $59.4_{\pm 3.4}$ |
| herbert-large-cased | $81.5_{\pm 2.5}$ | $80.7_{\pm 2.0}$ | $81.1_{\pm 1.5}$ | $\mathbf{82.4}_{\pm 1.1}$ | $77.6_{\pm 1.4}$ | $76.2_{\pm 2.7}$ | $\mathbf{76.7}_{\pm 2.1}$ |
| polish-roberta-base-v2 | $66.4_{\pm 3.0}$ | $\mathbf{86.5}_{\pm 3.9}$ | $75.1_{\pm 2.1}$ | $75.4_{\pm 1.5}$ | $69.7_{\pm 2.3}$ | $67.2_{\pm 1.9}$ | $66.9_{\pm 2.0}$ |
| polish-roberta-large-v2 | $\mathbf{84.6}_{\pm 3.6}$ | $77.5_{\pm 6.0}$ | $80.7_{\pm 2.9}$ | $81.7_{\pm 1.2}$ | $\mathbf{78.5}_{\pm 0.7}$ | $74.3_{\pm 3.7}$ | $75.8_{\pm 2.5}$ |
| **LLMs** | | | | | | | |
| deepseek-v3 zero-shot | $48.1_{\pm 0.3}$ | $90.6_{\pm 1.2}$ | $62.9_{\pm 0.6}$ | $49.5_{\pm 0.4}$ | $49.6_{\pm 0.2}$ | $47.9_{\pm 0.4}$ | $42.7_{\pm 0.5}$ |
| deepseek-v3 few-shot | $61.9_{\pm 0.3}$ | $96.1_{\pm 0.3}$ | $75.3_{\pm 0.1}$ | $59.0_{\pm 0.2}$ | $61.1_{\pm 0.4}$ | $63.7_{\pm 0.4}$ | $55.9_{\pm 0.3}$ |
| deepseek-v3 zero-shot+inst. | $84.7_{\pm 1.3}$ | $80.6_{\pm 0.7}$ | $\mathbf{82.6}_{\pm 0.6}$ | $70.7_{\pm 0.4}$ | $70.4_{\pm 0.6}$ | $74.8_{\pm 0.5}$ | $68.7_{\pm 0.4}$ |
| deepseek-v3 few-shot+inst. | $79.7_{\pm 0.9}$ | $82.0_{\pm 0.8}$ | $80.9_{\pm 0.9}$ | $68.4_{\pm 0.8}$ | $70.1_{\pm 0.6}$ | $76.4_{\pm 0.8}$ | $67.4_{\pm 0.8}$ |
| gpt-4o zero-shot | $42.8_{\pm 0.2}$ | $\mathbf{100.0}_{\pm 0.0}$ | $60.0_{\pm 0.2}$ | $47.6_{\pm 0.3}$ | $49.8_{\pm 0.2}$ | $46.8_{\pm 0.3}$ | $38.8_{\pm 0.3}$ |
| gpt-4o few-shot | $60.3_{\pm 0.2}$ | $98.8_{\pm 0.3}$ | $74.9_{\pm 0.3}$ | $57.5_{\pm 0.2}$ | $62.1_{\pm 0.1}$ | $66.5_{\pm 0.3}$ | $55.5_{\pm 0.3}$ |
| gpt-4o zero-shot+inst. | $85.7_{\pm 0.4}$ | $76.7_{\pm 0.8}$ | $80.9_{\pm 0.6}$ | $\mathbf{75.0}_{\pm 0.2}$ | $\mathbf{73.4}_{\pm 0.2}$ | $\mathbf{79.0}_{\pm 0.3}$ | $\mathbf{72.5}_{\pm 0.2}$ |
| gpt-4o few-shot+inst. | $\mathbf{86.0}_{\pm 1.9}$ | $75.1_{\pm 0.7}$ | $80.1_{\pm 0.6}$ | $68.5_{\pm 0.3}$ | $72.3_{\pm 0.5}$ | $76.5_{\pm 0.2}$ | $67.7_{\pm 0.3}$ |

Table 2: Average scores with standard deviation for all evaluated methods. The Precision, Recall, and F1 metrics were calculated considering only the dual quality class; the other metrics were for all classes, with 'm' as the macro average. Bold values indicate the highest scores for the type of method, and blue highlights the highest scores for each metric.

(Reimers and Gurevych, 2019), three e5 models (Wang et al., 2024) and mGTE (Zhang et al., 2024). Additionally, we choose four sentence-transformer models dedicated to the Polish language: st-polish-paraphrase-from-mpnet, st-polish-paraphrase-from-distilroberta (Dadas et al., 2024b) and two mmlw models (Dadas et al., 2024a).

**Transformer-based encoders** involves training pre-trained language model with classification head on top (a linear layer on top of the pooled output). We included evaluations of multilingual BERT (mBERT) (Devlin et al., 2019), multilingual XLM-RoBERTa (Conneau et al., 2020), and models specifically trained for Polish, such as HerBERT (Mroczkowski et al., 2021) and Polish RoBERTa (Dadas et al., 2020).

**LLMs** Advanced frontier models such as DeepSeek (DeepSeek-AI et al., 2025, 2024) and GPT-4o (OpenAI et al., 2024) were selected to evaluate how effectively cutting-edge LLMs handle dual quality review detection tasks under different prompting scenarios, including zero-shot and few-shot configurations, both with and without additional instruction (see more details about used prompts in Table 9).

### 4.3 Main Results

The experimental results from Table 2 clearly indicate notable differences among the three groups of tested models. Sentence-transformer models using SetFit generally achieved moderate precision scores (around 70-77%), suggesting that compressing sentence semantics into a single vector might result in information loss or inadequate semantic representation. Transformer-based encoders, particularly the larger, language-specific models such as polish-roberta-large-v2 (84.6%) and herbert-large-cased (81.5%), exhibited significantly stronger performance, comparable even with state-of-the-art conversational large language models (LLMs). Among LLMs, instructive prompting strategies (providing clear definitions of classes without explicit examples) improved performance, with the best precision results of 86% and 85.7% achieved by GPT-4o models with and without examples, respectively. It should be noted that the GPT-4o model with zero-shot instr. prompt achieved very good results for other measures as well. Interestingly, explicit few-shot examples sometimes distort the models and reduce detection efficiency overall. This may suggest that the chosen examples may not be representative and therefore helpful.

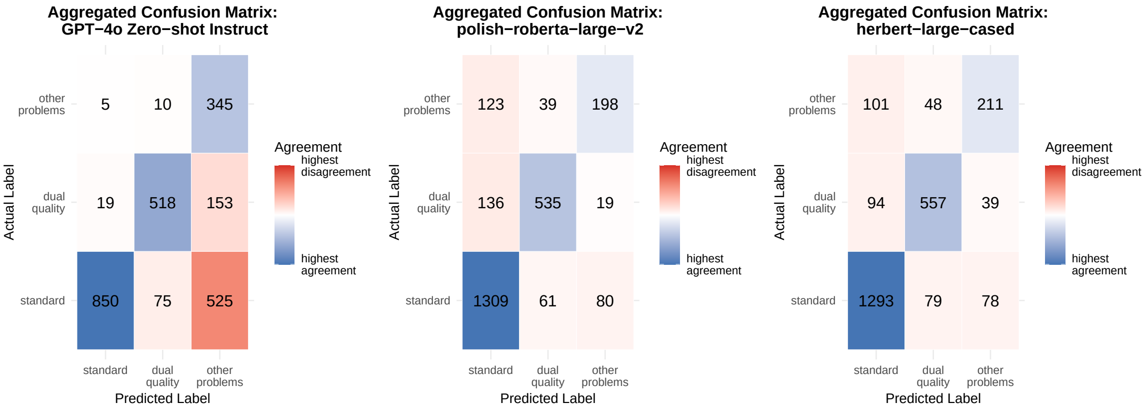


Figure 3: Confusion matrices aggregated from five experiments for selected models.

### 4.4 Errors Analysis

We conducted a detailed error analysis for selected models using classification confusion matrices visualized through heat maps. Specifically, we selected three representative models: GPT-4o (zero-shot+inst.), polish-roberta-large-v2 and herbert-large-cased. Figure 3 shows that the GPT-4o model exhibits substantial confusion between standard and 'other problems' reviews, while errors between standard and dual quality are less frequent. The polish-roberta-large-v2 model frequently identifies the standard reviews, achieving high accuracy for this category, but often misclassifies dual quality opinions as standard. Model herbert-large-cased often recognizes the dual quality reviews, achieving a high detection rate but also producing the most false positives for this class. Additional comparative analyses are presented in Figure 7 and Figure 8.

### 4.5 Robustness

As an additional experiment, we verified robustness of selected models, i.e., whether a slight change in the text, which does not significantly affect its meaning, can change the model's decision. We generated five additional test sets, which resulted from modifications to the original test set. The modifications are described in Table 3. We tested three selected models, the results are shown in Table 4. The percentage of differences in predictions was between 2.6 and 5.0. More often, larger text modifications like pl_chars influenced the change in decision.

### 4.6 Multilingual Transfer

To verify generalizability across markets and languages, we also explored multilingual transfer capabilities of our solution. For this purpose, we created a multilingual subset of reviews in English, German, and French (200,000 reviews for each language) selected from the AMAZON (Keung et al., 2020) dataset and our demo system. Next, we trained SetFit with paraphrase-multilingual-mpnet-base-v2[9] on the DQ dataset, and applied it to these reviews. Then we selected 500 AMAZON reviews and 200 reviews from demo system with the high-

| Name | Description |
|---|---|
| period | Remove (if present) or add (if absent) a period at the end of the review. |
| first_letter | Change the capitalization of the first letter of the first word in the review. If the first word is written in uppercase, change it to lowercase. |
| lower | Change text of the review to lowercase. |
| pl_chars | Replace the Polish characters *ą, ę, ć, ł, ń, ó, ż, ź* with their corresponding Latin alphabet characters, i.e., *a, e, c, l, n, o, z*. |
| pl_chars_once | The operation is the same as pl_chars, except that each letter can be changed once. |

Table 3: Descriptions of modifications applied to the test set for robustness verification.

| Modification | gpt-4o | polish-roberta | herbert |
|---|---|---|---|
| period | $4.0_{\pm 0.0}$ | $4.2_{\pm 1.0}$ | $5.0_{\pm 0.9}$ |
| first_letter | $4.0_{\pm 0.0}$ | $2.8_{\pm 0.7}$ | $2.6_{\pm 0.8}$ |
| lower | $5.0_{\pm 0.0}$ | $4.6_{\pm 0.5}$ | $4.2_{\pm 0.7}$ |
| pl_chars | $5.0_{\pm 0.0}$ | $4.6_{\pm 1.2}$ | $4.6_{\pm 0.8}$ |
| pl_chars_once | $4.0_{\pm 0.0}$ | $4.0_{\pm 1.4}$ | $3.6_{\pm 0.8}$ |

Table 4: Robustness verification results for GPT-4o (zero-shot+inst.), polish-roberta-large-v2 and herbert-large-cased. The values are the average and standard deviation of the model's decision disagreement for the original and modified reviews. To ensure consistent behavior in the GPT-4o model, we set the temperature to 0.0, resulting in a standard deviation of 0.0 across runs.

[9] One of the top multilingual sentence transformer at that time (2023).

| | Dual Quality class | | | All classes | | | |
|---|---|---|---|---|---|---|---|
| **Method** | **Precision** | **Recall** | **F1** | **Accuracy** | **mPrecision** | **mRecall** | **mF1** |
| **Transformer-based encoders** | | | | | | | |
| xlm-roberta-base | $69.5_{\pm 2.3}$ | $\mathbf{66.9_{\pm 6.8}}$ | $67.9_{\pm 2.9}$ | $\mathbf{73.0_{\pm 1.0}}$ | $55.5_{\pm 1.1}$ | $55.1_{\pm 2.1}$ | $55.0_{\pm 1.7}$ |
| xlm-roberta-large | $\mathbf{84.8_{\pm 3.8}}$ | $63.1_{\pm 4.8}$ | $\mathbf{72.3_{\pm 4.0}}$ | $72.6_{\pm 2.7}$ | $\mathbf{60.1_{\pm 2.7}}$ | $\mathbf{56.7_{\pm 3.9}}$ | $\mathbf{57.5_{\pm 3.3}}$ |
| **LLMs** | | | | | | | |
| deepseek-v3 zero-shot+inst. | $85.9_{\pm 1.8}$ | $52.3_{\pm 0.8}$ | $65.0_{\pm 0.3}$ | $49.5_{\pm 0.7}$ | $63.4_{\pm 1.3}$ | $\mathbf{58.7_{\pm 1.0}}$ | $49.1_{\pm 0.7}$ |
| deepseek-v3 few-shot+inst. | $\mathbf{91.9_{\pm 4.8}}$ | $50.6_{\pm 0.8}$ | $65.2_{\pm 1.8}$ | $44.3_{\pm 0.9}$ | $\mathbf{65.6_{\pm 2.2}}$ | $56.2_{\pm 1.2}$ | $46.1_{\pm 1.0}$ |
| gpt-4o zero-shot+inst. | $85.3_{\pm 1.3}$ | $46.6_{\pm 0.0}$ | $60.2_{\pm 0.3}$ | $\mathbf{52.6_{\pm 0.6}}$ | $62.3_{\pm 0.3}$ | $57.1_{\pm 0.3}$ | $\mathbf{49.6_{\pm 0.3}}$ |
| gpt-4o few-shot+inst. | $80.2_{\pm 1.1}$ | $46.6_{\pm 0.0}$ | $58.9_{\pm 0.3}$ | $41.6_{\pm 0.6}$ | $61.4_{\pm 0.5}$ | $50.2_{\pm 1.0}$ | $42.7_{\pm 0.5}$ |

Table 5: Evaluation results for selected models on a multilingual dataset.

est dual quality scores. Manual verification showed that most were actually standard, so we randomly limited standard reviews to 130, yielding **206** final examples (**58** dual quality, **18** other problems, **130** standard). The dataset thus prepared was used as a multilingual test set. We conducted an experiment in which we tested methods based on multilingual models trained as in Section 4.1 on the Polish training subset or, in the case of LLMs, using the same prompts. The results for the selected models are presented in Table 5. Considering the precision of the classifier, the highest score was achieved by the DeepSeek-V3 (91.9%) model, interestingly in this case, adding examples to the instructions in the prompt gave a higher score. Of the group of transformer-based encoders, the highest score was achieved by xlm-roberta-large (84.8%). Although the difference in performance on the basis of precision is significant, it is important to note the low values of the recall measure for LLMs, compared to encoders. All results for this experiment are available in Table 11.

## 5 Deployment and Practical Considerations

During the evaluation, a key objective was to achieve high precision, thereby minimizing the number of false positive recommendations. Since each flagged instance undergoes final verification by a human analyst, the primary goal is to reduce the analyst's workload by minimizing the number of irrelevant alerts. This approach accepts the possibility of missing some true dual quality cases (i.e., allowing for a certain level of false negatives) in favor of ensuring that the identified cases are highly likely to be accurate. A product with several dual quality reviews will be selected for further analysis to verify whether this issue genuinely exists in its case.

The proposed solution is implemented as a standalone service within a local infrastructure and is exclusively dedicated to UOKiK employees (Poland's Office of Competition and Consumer Protection). The system is currently not accessible to the public or external users. Although the system can analyze multilingual content, the current deployment prioritizes support for the Polish language to align with the context of Polish consumers and UOKiK's mandate within the Polish market.

Given the results of the evaluation and the above assumptions, we would recommend using the polish-robert-large-v2 model for a production deployment. Selecting the locally deployable model presents a pragmatic and efficient choice, particularly when minimizing external dependencies and ensuring consistent, low-latency inference. It should be noted that this language-specific component is modular; for deployment within other European consumer protection agencies analogous to UOKiK, the model could be readily substituted with an equivalent model fine-tuned for the respective national language (e.g., a German BERT for a German institution) or multilingual model like XLM-RoBERTa.

## 6 Conclusion

In this work, we presented the entire process of preparing a solution for detecting the problem of dual quality based on product reviews. Our three key findings are: First, mentions of dual quality in product reviews are rare, in our case appearing only a few hundred times. Second, smaller language-specific transformer-based encoders finetuned for the task perform comparably to larger LLMs. Finally, including examples in prompts for LLMs can degrade performance compared to using only task-specific instructions.

## Acknowledgments

Project co-financed/financed by the National Centre for Research and Development (https://www.gov.pl/web/ncbr-en) under the programme Infostrateg III.

## References


Noa Avigdor, Guy Horowitz, Ariel Raviv, and Stav Yanovsky Daye. 2023. Consistent text categorization using data augmentation in e-commerce. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 313–321, Toronto, Canada. Association for Computational Linguistics.

Lucia Bartkova and Mária Sirotiaková. 2021. Dual quality and its influence on consumer behaviour according to the income. *SHS Web of Conferences*, 92.

Lucia Bartkova and Lenka Veselovska. 2023. Does dual quality of products in the european union truly bother consumers? *Marketing and Management of Innovations*, 14.

Lucia Bartkova, Lenka Veselovska, Marianna Sramkova, and Jan Zavadsky. 2021. Dual quality of products: myths and facts through the opinions of millennial consumers. *Marketing and Management of Innovations*.

L. Bartková and L. Veselovská. 2024. Consumer behaviour under dual quality of products: Does testing reveal what consumers experience? *IIMB Management Review*, 36:171–184.

Lucia Bartková. 2019. How do consumers perceive the dual quality of goods and its economic aspects in the european union? an empirical study. *Problems and Perspectives in Management*, 17.

Lucia Bartková, Lenka Veselovská, and Katarína Zimermanová. 2018. Possible solutions to dual quality of products in the european union. *Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration*, 26.

I. Botunac, M. Brkić Bakarić, and M. Matetić. 2024. Comparing fine-tuning and prompt engineering for multi-class classification in hospitality review analysis. *Applied Sciences (Switzerland)*, 14.

Chambers. Dual Quality of Food Products. `https://chambers.com/legal-trends/dual-quality-of-food-products`. [Online; accessed 06-March-2025].

European Commission. 2018. Dual quality of food: European Commission releases common testing methodology. `https://ec.europa.eu/commission/presscorner/detail/en/ip_18_4122`. [Online; accessed 06-March-2025].

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Pre-training polish transformer-based language models at scale. In *Artificial Intelligence and Soft Computing*, pages 301–314. Springer International Publishing.

Slawomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2024a. PIRB: A comprehensive benchmark of Polish dense and hybrid text retrieval methods. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12761–12774, Torino, Italia. ELRA and ICCL.

Sławomir Dadas, Marek Kozłowski, Rafał Poświata, Michał Perełkiewicz, Marcin Białas, and Małgorzata Grębowiec. 2024b. A support system for the detection of abusive clauses in b2c contracts. *Artificial Intelligence and Law*.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

EU Monitor. The better enforcement and modernisation of Union consumer protection rules. `https://www.eumonitor.eu/9353000/1/j4nvhdfcs8bljza_j9vvik7m1c3gyxp/vme85bbfssxo`. [Online; accessed 06-March-2025].

Joint Research Centre European Commission. 2023. Same pack, different ingredients: Is dual quality down-branded in EU food? `https://joint-research-centre.ec.europa.eu/jrc-news-and-updates/same-pack-different-ingredients-dual-quality-down-branded-eu-food-2023-07-24_en`. [Online; accessed 06-March-2025].

European Parliamentary Research Service (EPRS) European Parliament. 2017. European Commission guidelines on dual quality of branded food products. `https://www.europarl.europa.eu/RegData/etudes/BRIE/2017/608804/EPRS_BRI%282017%29608804_EN.pdf`. [PDF; accessed 06-March-2025].

European Parliamentary Research Service (EPRS) European Parliament. 2019. Dual quality of products – State of play. `https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/644192/EPRS_BRI(2019)644192_EN.pdf`. [Online; accessed 06-March-2025].

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Gilad Fuchs, Ido Ben-shaul, and Matan Mandelbrod. 2022. Is it out yet? automatic future product releases extraction from web data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 263–271, Abu Dhabi, UAE. Association for Computational Linguistics.

Shansan Gong, Zelin Zhou, Shuo Wang, Fengjiao Chen, Xiujie Song, Xuezhi Cao, Yunsen Xian, and Kenny Zhu. 2023. Transferable and efficient: Unifying dynamic multi-domain product categorization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 476–486, Toronto, Canada. Association for Computational Linguistics.

Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over past, evolve for future: Forecasting temporal trends for fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 116–125, Toronto, Canada. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Y. Mamani-Coaquira and E. Villanueva. 2024. A review on text sentiment analysis with machine learning and deep learning techniques. *IEEE Access*, 12:193115–193130.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Ravindra Nayak and Nikesh Garera. 2022. Deploying unified BERT moderation model for E-commerce reviews. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 540–547, Abu Dhabi, UAE. Association for Computational Linguistics.

Stephen Obadinma, Faiza Khan Khattak, Shirley Wang, Tania Sidhorn, Elaine Lau, Sean Robertson, Jingcheng Niu, Winnie Au, Alif Munim, and Karthik Raja Kalaiselvi Bhaskar. 2022. Bringing the state-of-the-art to customers: A neural agent assistant framework for customer service support. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 440–450, Abu Dhabi, UAE. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Soham Parikh, Mitul Tiwari, Prashil Tumbade, and Quaizar Vohra. 2023. Exploring zero and few-shot techniques for intent classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 744–751, Toronto, Canada. Association for Computational Linguistics.

Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. 2024. PL-MTEB: Polish Massive Text Embedding Benchmark. *Preprint*, arXiv:2405.10138.

Nils Reimers and Iryna Gurevych. 2019. 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.

Safe Food Advocacy Europe (SAFE). Dual Food Quality Project. `https://www.safefoodadvocacy.eu/projects/dual-food-quality-project/`. [Online; accessed 06-March-2025].

J. Satjathanakul and T. Siriborvornratanakul. 2024. Sentiment analysis in product reviews in thai language. *International Journal of Information Technology (Singapore)*.

Xiaoyu Shen, Akari Asai, Bill Byrne, and Adria De Gispert. 2023. xPQA: Cross-lingual product question answering in 12 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 103–115, Toronto, Canada. Association for Computational Linguistics.

Hasan Tercan and Tobias Meisen. 2022. Machine learning and deep learning based predictive quality in manufacturing: a systematic review.

The European Consumer Organisation (BEUC). 2018. Dual product quality across Europe: state-of-play and the way forward. https://www.beuc.eu/sites/default/files/publications/beuc-x-2018-031_beuc_position_paper_on_dual_quality.pdf. [Online; accessed 06-March-2025].

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint*.

Lenka Veselovská. 2022. Dual quality of products in europe: a serious problem or a marketing opportunity? *Total Quality Management and Business Excellence*, 33.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Tianqi Wang, Lei Chen, Xiaodan Zhu, Younghun Lee, and Jing Gao. 2023. Weighted contrastive learning with false negative control to help long-tailed product classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 574–580, Toronto, Canada. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Ján Závadský and Vladimír Hiadlovský. 2020. Economic problems of dual quality of everyday consumer goods. *Economic Annals-XXI*, 185.

# A Dual Qulity Regulations

The regulatory response to dual quality has evolved significantly within the European Union. The European Commission's 2017 guidelines clarified that while product differentiation is not inherently illegal, misleading consumers violates EU consumer protection laws (European Parliament, 2017, 2019). The Commission's Joint Research Centre (JRC) introduced a harmonized testing methodology to assess product composition variations (Commission, 2018; European Commission, 2023) systematically. Additionally, the Omnibus Directive amended Directive 2005/29/EC, classifying dual quality marketing as misleading when substantial differences exist without a legitimate justification (Chambers). These measures aim to enhance market transparency and prevent unfair commercial practices. However, challenges remain in enforcement and uniform interpretation across Member States (EU Monitor). Recent research shows that while the prevalence of dual quality food products declined from 31% in 2018 to 24% in 2021, concerns persist regarding non-food items, as similar discrepancies have been identified in household and personal care products (European Commission, 2023).

Furthermore, consumer advocacy organizations such as BEUC argue that enforcement mechanisms must be strengthened to ensure compliance across all product categories (The European Consumer Organisation (BEUC), 2018). The SAFE initiative also supports enhanced consumer education and reporting mechanisms to empower individuals to identify and challenge dual quality practices (Safe Food Advocacy Europe (SAFE)). These ongoing legal and regulatory efforts underscore the EU's commitment to fair competition and consumer protection, yet continued vigilance and adaptation of enforcement strategies remain necessary.

# B DQ Dataset Details

## B.1 Annotation Process Details

We established a structured data labelling policy to annotate the data, i.e., assign each opinion or review to its appropriate category. This policy provides clear classification criteria for opinions categorized as *dual quality*, *other problems*, or *standard* (see Table 6 for detailed definitions). The annotation process followed predefined guidelines to ensure consistency and reliability, and where necessary, ambiguous cases were resolved through annotators' review.

Examples of labeled reviews from the DQ database, annotated according to the established data annotation protocol and accompanied by annotator comments, are presented in Table 7.

| Label | Description |
|---|---|
| dual quality | The review contains information about the fact that the customer bought the same product in two countries and noticed a difference in quality, performance, composition, etc. It is not necessary to give the exact names of the countries, phrases such as "abroad" or "in our country" are sufficient. The customer is comparing two same products or groups of products. Indicating a difference in price, availability or using a general statement such as "there are differences between products purchased in France and Poland" are **NOT** classified as dual quality, but as standard review. |
| other problems | The review does not identify the problem of dual quality, but provides information about other problems, among which we can distinguish:<br>– differences in products due to a different place of purchase (same market), place of packaging or batch received,<br>– problems with the product itself that require deeper analysis e.g., deterioration over time,<br>– practices that are illegal and/or violate customer rights e.g., the product is probably counterfeit, suspected fraud, misleading the customer, no instructions in the required language, no expiration date, etc.. |
| standard | A standard product review in which the comments described are about the product itself and do not indicate problems addressed by the labels "dual quality" or "other problems". |

Table 6: Annotation Guidelines.

### B.2 Other Problems Identified in Products or Services

When labeling the data, annotators identified opinions explicitly reflecting dual quality issues and comments pointing to specific problems related to services or products. These additional insights enabled deeper exploration and facilitated the creation of a comprehensive taxonomy of consumer issues. Figure 5 demonstrates that more than half of the reported problems concern probable counterfeit products, differences dependent on the place of purchase within the same market, quality deterioration over time, mismatches between received products and orders, misleading information, suspicions of fraud, and variations related to packaging, batch, or package size. Recognizing and categorizing these issues may be crucial for targeted interventions and regulatory measures to strengthen consumer trust and improve market standards beyond dual quality considerations alone.

## C Experiments Details

**Baseline** For the baseline model, the text was first lemmatized. Then the following phrases were searched: `anglia, angielski, szkocja, szkocki, irlandia, irlandzki, walia, walijski, dania, duński, finlandia, fiński, norwegia, norweski, szwecja, szwedzki, szwajcaria, szwajcarski, estonia, estoński, łotwa, łotewski, litwa, litewski, austria, austryjacki, belgia, belgijski, francja, francuski, niemcy, niemiecki, włochy, włoski, holandia, niderlandzki, holenderski, usa, kanada, kanadyjski, meksyk, meksykański, ukraina, ukraiński, rosja, rosyjski, białoruś, białoruski, polska, polski, czechy, czeski, słowacja, słowacki, węgry, węgierski, rumunia, rumuński, bułgaria, bułgarski, grecja, grecki, hiszpania, hiszpański, brazylia, brazylijski, portugalia, portugalski, australia, australijski, nowa zelandia, maoryjski, gruzja, gruziński, izrael, hebrajski, egipt, arabski, turcja, turecki, chiny, chiński, korea, koreański, japonia, japoński, indie, hinduski.`
If one or more of the above phrases were found, the review was classified as dual quality.

**SetFit + sentence transformer** During training, we used the following hyperparameters: learning rate=2e-5 (same for sentence transformer fine-tuning and logistic regression classifier), batch size=8, epochs=1, number of iterations for contrastive=1. We adopted AdamW optimizer.

**Transformer-based encoders** During training, we used the following hyperparameters: learning rate=2e-6, batch size=8, epochs=10. We adopted AdamW optimizer.

**LLMs** The models were evaluated using APIs. For the main experiments the temperature was set to 0.1, for robustness verification to guarantee determinism it was reduced to 0.0. The prompts used are shown in Table 9.

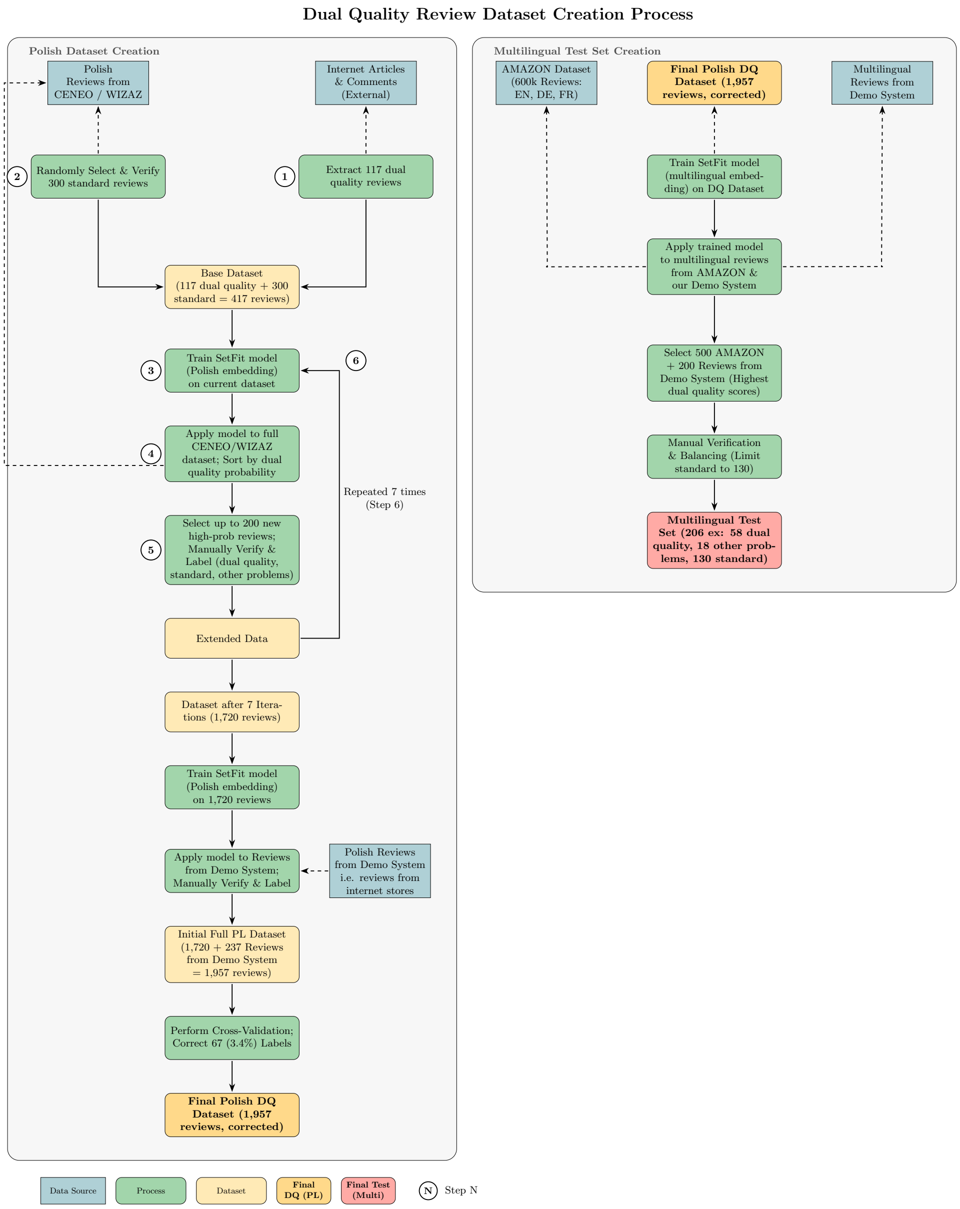


Figure 4: Diagram showing the process of preparing DQ and multilingual datasets.

| Original review text | Translated review text | Label | Additional Comment |
|---|---|---|---|
| Fantastyczny zapach i produkt z chemii niemieckiej, więc o wiele bardziej intensywny niż te, produkowane na polski rynek. | Fantastic fragrance and a product of German chemistry, so much more intense than those made for the Polish market. | dual quality | - |
| Jedna z moich ulubionych kaw, zwłaszcza ta w wersji z Niemiec. O wiele bardziej aromatyczna niż proponowana na rynek Polski | One of my favorite coffees, especially the version from Germany. Much more aromatic than the one offered on the Polish market. | dual quality | - |
| poprzedni model Beko kupiony 9 lat temu był lepszy | The previous Beko model bought 9 years ago was better. | other problems | deterioration in quality over time |
| Tester w drogerii(w centrum handlowym) był dużo bardziej trwały i intensywniejszy niż ten kupiony przez internet. Zastanawiające. | The tester in the drugstore (at the shopping mall) was much more long-lasting and intense than the one purchased online. Intriguing. | other problems | difference depending on the place of purchase (same market) |
| Maska spełnia swoje zadanie. Rewelacyjnie pachnie. | The mask does its job. It smells amazing. | standard | - |
| soczewki produkowane poza Europą mają kiepską jakość | Lenses produced outside Europe are of poor quality. | standard | general statement |

Table 7: A list of samples from DQ dataset. The original text of the review was translated into English using GPT-4o.

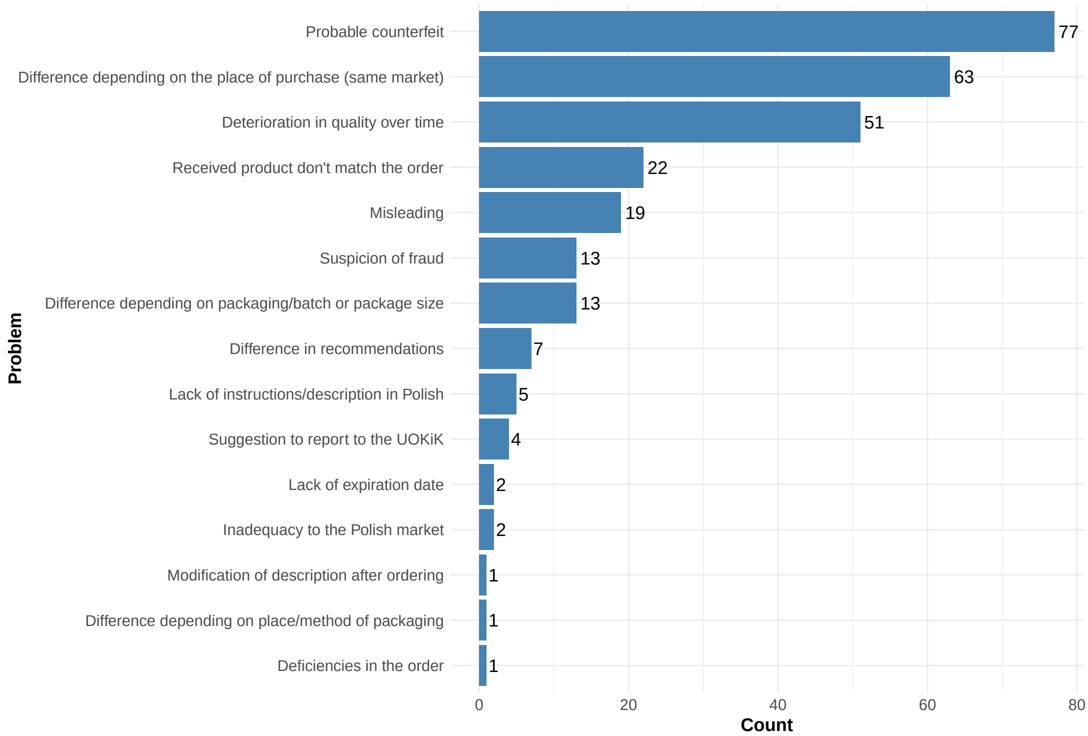


Figure 5: Taxonomy of different product or service issues recognized in reviews.

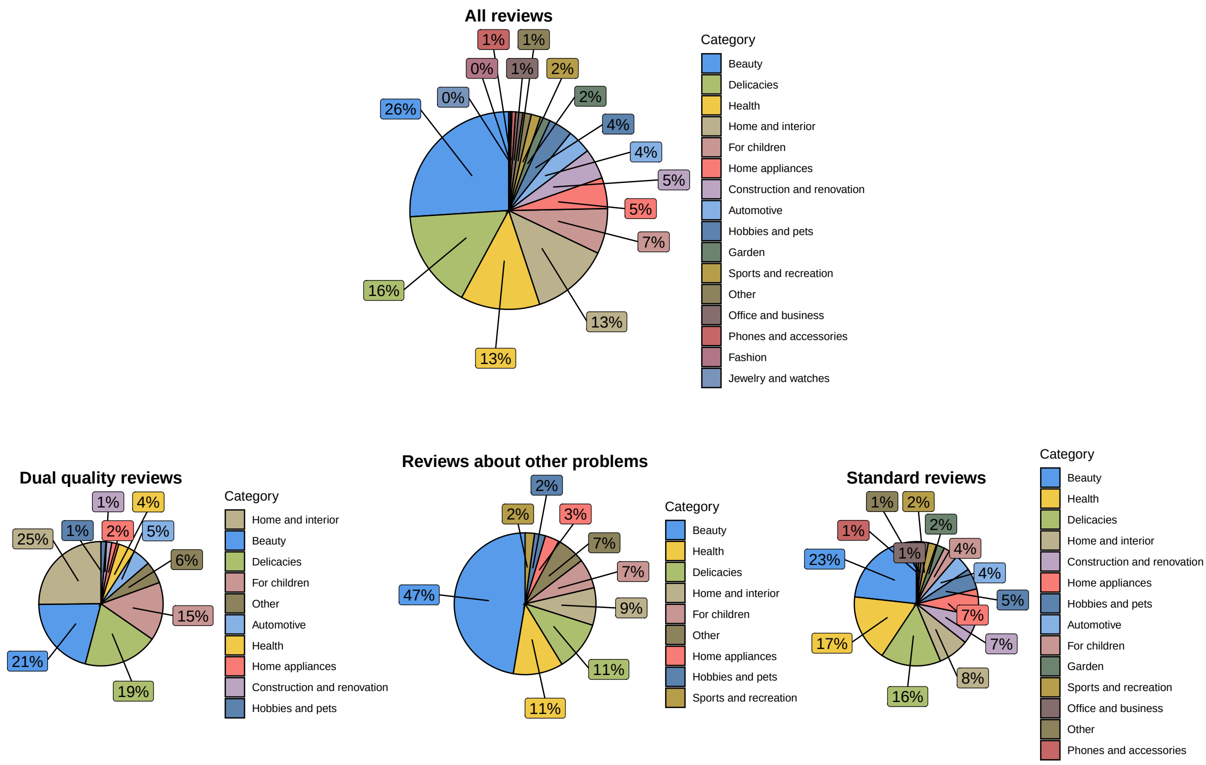


Figure 6: Charts illustrating (1) all product reviews categorized by product type (top) and (2) the distribution of product categories across various types of reviews (bottom).

| Name in Paper | HF Name |
|---|---|
| LaBSE | sentence-transformers/LaBSE |
| para-multi-mpnet-base-v2 | sentence-transformers/paraphrase-multilingual-mpnet-base-v2 |
| para-multi-MiniLM-L12-v2 | sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 |
| multi-e5-small | intfloat/multilingual-e5-small |
| multi-e5-base | intfloat/multilingual-e5-base |
| multi-e5-large | intfloat/multilingual-e5-large |
| gte-multi-base | Alibaba-NLP/gte-multilingual-base |
| st-polish-para-mpnet | sdadas/st-polish-paraphrase-from-mpnet |
| st-polish-para-distilroberta | sdadas/st-polish-paraphrase-from-distilroberta |
| mmlw-roberta-base | sdadas/mmlw-roberta-base |
| mmlw-roberta-large | sdadas/mmlw-roberta-large |
| mBERT | google-bert/bert-base-multilingual-cased |
| xlm-roberta-base | FacebookAI/xlm-roberta-base |
| xlm-roberta-large | FacebookAI/xlm-roberta-large |
| herbert-base-cased | allegro/herbert-base-cased |
| herbert-large-cased | allegro/herbert-large-cased |
| polish-roberta-base-v2 | sdadas/polish-roberta-base-v2 |
| polish-roberta-large-v2 | sdadas/polish-roberta-large-v2 |
| deepseek-v3* | deepseek-ai/DeepSeek-V3 |
| gpt-4o* | - |

Table 8: Model names as referenced in the paper, and corresponding Hugging Face Hub identifiers. An asterisk (*) indicates models accessed via REST APIs: DeepSeek-V3 (`https://api-docs.deepseek.com/`) and GPT-4o (`https://platform.openai.com/docs/api-reference/introduction`).

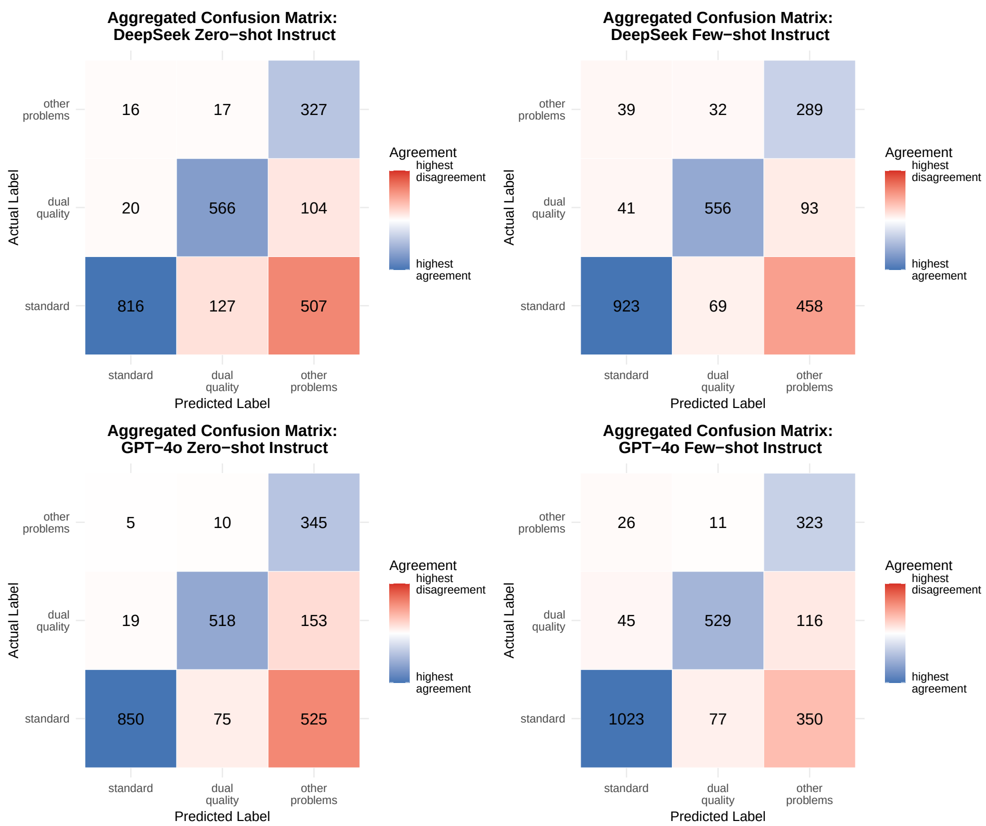


Figure 7: Confusion matrices aggregated from five experiments for DeepSeek and GPT-4o models in zero-shot and few-shot instruction-based configurations.

| Type | Prompt |
|---|---|
| zero-shot | Przypisz podaną niżej opinie do jednej z trzech klas: "dual quality", "other problems" lub "standard".<br>W odpowiedzi podaj jedynie nazwę klasy, bez dodatkowego komentarza.<br>Treść opinii:<br><*review*> |
| few-shot | Przypisz podaną niżej opinie do jednej z trzech klas: "dual quality", "other problems" lub "standard".<br><br>**Przykłady:**<br>Kapsułki są lepsze, niż na polski rynek tej samej firmy. – dual quality<br>Dobry smak kawy. Kraj pochodzenia Niemcy. Nie jest tak kwaśna jak kupiona w kraju. – dual quality<br>Mój ulubiony zapach. Sądzę jednak, że są dużo mniej trwałe niż te, które poprzednim razem kupiłam w sephorze. – other problems<br>Proszek może i z Niemiec, ale produkcja Czechy - wprowadzanie klienta w błąd. – other problems<br>Niezły preparat. Łagodzi trochę bóle i zmęczenie oczu. Stosuję od czasu do czasu. – standard<br>jest ok, nie zauważyłam większej różnicy między "polską" a "niemiecką" wersją – standard<br><br>W odpowiedzi podaj jedynie nazwę klasy, bez dodatkowego komentarza.<br>Treść opinii:<br><*review*> |
| zero-shot+inst. | Przypisz podaną niżej opinie do jednej z trzech klas: "dual quality", "other problems" lub "standard".<br><br>**Wytyczne dla każdej z klas:**<br>**"dual quality" (podwójna jakość)** – opinia zawiera informacje o tym, że klient kupił ten sam produkt w dwóch krajach i zauważył różnicę w jakości, wydajności, składzie itp. Nie jest konieczne podawanie dokładnych nazw krajów, wystarczą zwroty takie jak „za granicą" lub „w naszym kraju". Klient porównuje dwa takie same produkty lub grupy produktów. Wskazanie różnicy w cenie, dostępności lub ogólne stwierdzenie, takie jak „istnieją różnice między produktami zakupionymi we Francji i w Polsce" nie są klasyfikowane jako podwójna jakość.<br><br>**"other problems" (inne problemy)** – opinia nie wskazuje na problem podwójnej jakości, ale dostarcza informacji o innych problemach, wśród których możemy wyróżnić: różnice w produktach wynikające z innego miejsca zakupu (ten sam rynek), miejsca pakowania lub otrzymanej partii; problemy z samym produktem wymagające głębszej analizy np. pogorszenie jakości z upływem czasu; praktyki niezgodne z prawem i/lub naruszające prawa klienta np. produkt jest prawdopodobnie podrobiony, podejrzenie oszustwa, wprowadzanie klienta w błąd, brak instrukcji w wymaganym języku, brak daty ważności itp.<br><br>**"standard"** – standardowa opinia o produkcie, w której opisane uwagi dotyczą samego produktu i nie wskazują na problemy omówione przy klasach „podwójna jakość" lub „inne problemy".<br><br>W odpowiedzi podaj jedynie nazwę klasy, bez dodatkowego komentarza.<br>Treść opinii:<br><*review*> |
| few-shot+inst. | Przypisz podaną niżej opinie do jednej z trzech klas: "dual quality", "other problems" lub "standard".<br><br>**Wytyczne dla każdej z klas:**<br>**"dual quality" (podwójna jakość)** – opinia zawiera informacje o tym, że klient kupił ten sam produkt w dwóch krajach i zauważył różnicę w jakości, wydajności, składzie itp. Nie jest konieczne podawanie dokładnych nazw krajów, wystarczą zwroty takie jak „za granicą" lub „w naszym kraju". Klient porównuje dwa takie same produkty lub grupy produktów. Wskazanie różnicy w cenie, dostępności lub ogólne stwierdzenie, takie jak „istnieją różnice między produktami zakupionymi we Francji i w Polsce" nie są klasyfikowane jako podwójna jakość.<br>**Przykłady:** "Kapsułki są lepsze, niż na polski rynek tej samej firmy.", "Dobry smak kawy. Kraj pochodzenia Niemcy. Nie jest tak kwaśna jak kupiona w kraju."<br><br>**"other problems" (inne problemy)** – opinia nie wskazuje na problem podwójnej jakości, ale dostarcza informacji o innych problemach, wśród których możemy wyróżnić: różnice w produktach wynikające z innego miejsca zakupu (ten sam rynek), miejsca pakowania lub otrzymanej partii; problemy z samym produktem wymagające głębszej analizy np. pogorszenie jakości z upływem czasu; praktyki niezgodne z prawem i/lub naruszające prawa klienta np. produkt jest prawdopodobnie podrobiony, podejrzenie oszustwa, wprowadzanie klienta w błąd, brak instrukcji w wymaganym języku, brak daty ważności itp.<br>**Przykłady:** "Mój ulubiony zapach. Sądzę jednak, że są dużo mniej trwałe niż te, które poprzednim razem kupiłam w sephorze", "Proszek może i z Niemiec, ale produkcja Czechy - wprowadzanie klienta w błąd."<br><br>**"standard"** – standardowa opinia o produkcie, w której opisane uwagi dotyczą samego produktu i nie wskazują na problemy omówione przy klasach „podwójna jakość" lub „inne problemy".<br>**Przykłady:** "Niezły preparat. Łagodzi trochę bóle i zmęczenie oczu. Stosuję od czasu do czasu.", "jest ok, nie zauważyłam większej różnicy między "polską" a "niemiecką" wersją"<br><br>W odpowiedzi podaj jedynie nazwę klasy, bez dodatkowego komentarza.<br>Treść opinii:<br><*review*> |

Table 9: Prompts used during LLMs evaluation. Bold text and blank lines were added only for readability of the table. For non-Polish speakers, translated prompts available in Table 10.

| Type | Prompt |
|---|---|
| zero-shot | Assign the following review to one of three classes: “dual quality”, “other problems” or “standard”.<br>In your answer, provide only the name of the class, without additional comment.<br>Review text:<br><*review*> |
| few-shot | Assign the following review to one of three classes: “dual quality”, “other problems” or “standard”.<br><br>**Examples:**<br>The capsules are better than those on the Polish market from the same company. – dual quality<br>Good coffee taste. Country of origin: Germany. It is not as acidic as the one bought in the country. – dual quality<br>My favorite scent. However, I think it’s much less long-lasting than the one I bought at Sephora last time. – other problems<br>The powder may be from Germany, but it’s made in the Czech Republic - misleading the customer. – other problems<br>Decent product. It slightly alleviates eye pain and fatigue. I use it occasionally. – standard<br>It’s okay, I didn’t notice much difference between the "Polish" and "German" version. – standard<br><br>In your answer, provide only the name of the class, without additional comment.<br>Review text:<br><*review*> |
| zero-shot+inst. | Assign the following review to one of three classes: “dual quality”, “other problems” or “standard”.<br><br>**Guidelines for each category:**<br>**"dual quality"** – The review includes information that the customer purchased the same product in two different countries and noticed a difference in quality, performance, composition, etc. It is not necessary to specify the exact names of the countries; phrases like "abroad" or "in our country" are sufficient. The customer compares two identical products or groups of products. Indicating a difference in price, availability, or a general statement such as "there are differences between products purchased in France and Poland" is not classified as dual quality.<br><br>**"other problems"** – The review does not indicate an issue of dual quality but provides information on other problems, which can include: differences in products resulting from a different place of purchase (same market), place of packaging, or the received batch; problems with the product itself requiring deeper analysis, such as deterioration in quality over time; practices that are illegal and/or violate customer rights, such as the product potentially being counterfeit, suspicion of fraud, misleading the customer, lack of instructions in the required language, lack of an expiration date, etc.<br><br>**"standard"** – A standard product review where the comments pertain only to the product itself and do not indicate the problems discussed in the "dual quality" or "other problems" categories.<br><br>In your answer, provide only the name of the class, without additional comment.<br>Review text:<br><*review*> |
| few-shot+inst. | Assign the following review to one of three classes: “dual quality”, “other problems” or “standard”.<br><br>**Guidelines for each category:**<br>**"dual quality"** – The review includes information that the customer purchased the same product in two different countries and noticed a difference in quality, performance, composition, etc. It is not necessary to specify the exact names of the countries; phrases like "abroad" or "in our country" are sufficient. The customer compares two identical products or groups of products. Indicating a difference in price, availability, or a general statement such as "there are differences between products purchased in France and Poland" is not classified as dual quality.<br>**Examples:** "The capsules are better than those on the Polish market from the same company.", "Good coffee taste. Country of origin: Germany. It is not as acidic as the one bought in the country."<br><br>**"other problems"** – The review does not indicate an issue of dual quality but provides information on other problems, which can include: differences in products resulting from a different place of purchase (same market), place of packaging, or the received batch; problems with the product itself requiring deeper analysis, such as deterioration in quality over time; practices that are illegal and/or violate customer rights, such as the product potentially being counterfeit, suspicion of fraud, misleading the customer, lack of instructions in the required language, lack of an expiration date, etc.<br>**Examples:** "My favorite scent. However, I think it’s much less long-lasting than the one I bought at Sephora last time.", "The powder may be from Germany, but it’s made in the Czech Republic - misleading the customer."<br><br>**"standard"** – A standard product review where the comments pertain only to the product itself and do not indicate the problems discussed in the "dual quality" or "other problems" categories.<br>**Examples:** "Decent product. It slightly alleviates eye pain and fatigue. I use it occasionally.", "It’s okay, I didn’t notice much difference between the "Polish" and "German" version."<br><br>In your answer, provide only the name of the class, without additional comment.<br>Review text:<br><*review*> |

Table 10: Translated prompts from Table 9 used during LLMs evaluation.

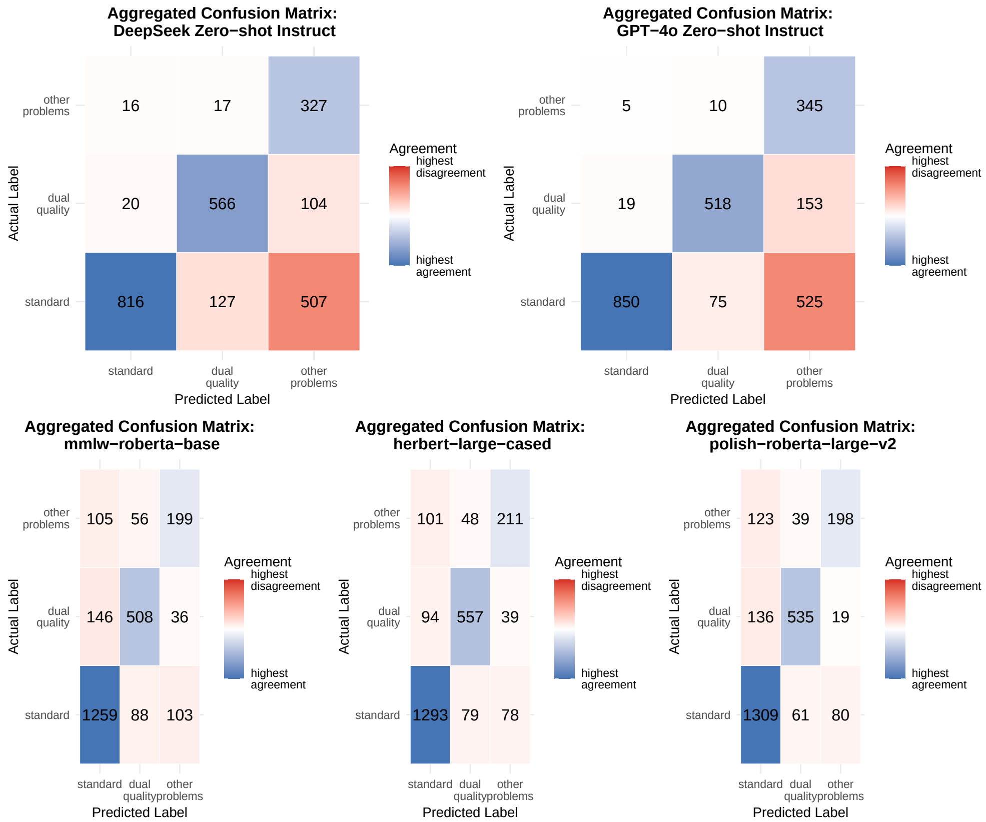


Figure 8: Confusion matrices aggregated from five experiments for best performing LLMs and top-performing local models.

| | Dual Quality class | | | All classes | | | |
|---|---|---|---|---|---|---|---|
| **Method** | **Precision** | **Recall** | **F1** | **Accuracy** | **mPrecision** | **mRecall** | **mF1** |
| **SetFit + sentence transformers** | | | | | | | |
| LaBSE | 74.0±8.7 | 37.9±12.1 | 49.1±11.4 | 70.1±3.6 | 55.5±3.6 | 47.6±4.1 | 48.4±4.8 |
| para-multi-mpnet-base-v2 | 69.4±3.4 | 45.5±4.4 | 54.8±3.0 | 67.1±1.8 | 53.2±1.4 | 49.6±0.8 | 50.2±0.6 |
| para-multi-MiniLM-L12-v2 | 69.3±3.2 | 40.3±7.8 | 50.7±7.4 | 62.9±2.0 | 49.3±1.9 | 43.2±2.7 | 44.6±3.0 |
| multi-e5-small | 74.0±4.1 | 41.7±4.8 | 53.2±4.2 | 72.9±1.3 | 49.1±1.5 | 46.2±1.5 | 45.5±1.7 |
| multi-e5-base | 78.4±4.8 | 45.9±19.1 | 54.6±19.1 | **73.4±4.1** | 54.1±5.0 | 49.2±7.1 | 48.6±9.1 |
| multi-e5-large | 0.0±0.0 | 0.0±0.0 | 0.0±0.0 | 63.1±0.0 | 21.0±0.0 | 33.3±0.0 | 25.8±0.0 |
| gte-multi-base | **81.7±4.9** | **58.0±4.7** | **67.7±3.7** | 71.6±3.3 | **57.2±2.7** | **52.5±2.6** | **54.0±2.7** |
| **Transformer-based encoders** | | | | | | | |
| mBERT | 61.7±19.5 | 6.6±4.3 | 11.1±6.7 | 62.1±2.8 | 43.8±6.3 | 34.7±2.2 | 30.3±3.3 |
| xlm-roberta-base | 69.5±2.3 | **66.9±6.8** | 67.9±2.9 | **73.0±1.0** | 55.5±1.1 | 55.1±2.1 | 55.0±1.7 |
| xlm-roberta-large | **84.8±3.8** | 63.1±4.8 | **72.3±4.0** | 72.6±2.7 | **60.1±2.7** | **56.7±3.9** | **57.5±3.3** |
| **LLMs** | | | | | | | |
| deepseek-v3 zero-shot | 47.6±1.9 | 86.2±2.8 | 61.4±2.3 | 32.4±0.9 | 46.9±1.5 | 39.4±1.0 | 28.8±0.9 |
| deepseek-v3 few-shot | 62.8±1.4 | 70.7±1.4 | **66.5±0.7** | 35.6±0.6 | 54.3±0.7 | 46.7±1.8 | 36.7±0.7 |
| deepseek-v3 zero-shot+inst. | 85.9±1.8 | 52.3±0.8 | 65.0±0.3 | 49.5±0.7 | 63.4±1.3 | **58.7±1.0** | 49.1±0.7 |
| deepseek-v3 few-shot+inst. | **91.9±4.8** | 50.6±0.8 | 65.2±1.8 | 44.3±0.9 | **65.6±2.2** | 56.2±1.2 | 46.1±1.0 |
| gpt-4o zero-shot | 38.8±0.6 | **86.8±2.2** | 53.6±1.0 | 33.3±0.6 | 47.4±0.2 | 36.8±0.7 | 27.0±0.4 |
| gpt-4o few-shot | 58.5±0.8 | 73.6±0.8 | 65.1±0.4 | 34.1±0.6 | 55.8±0.6 | 48.1±2.3 | 34.7±0.7 |
| gpt-4o zero-shot+inst. | 85.3±1.3 | 46.6±0.0 | 60.2±0.3 | **52.6±0.6** | 62.3±0.3 | 57.1±0.3 | **49.6±0.3** |
| gpt-4o few-shot+inst. | 80.2±1.1 | 46.6±0.0 | 58.9±0.3 | 41.6±0.6 | 61.4±0.5 | 50.2±1.0 | 42.7±0.5 |

Table 11: Evaluation results on a multilingual dataset consisting of English, German and French reviews. In red were marked results showing an example of when a multilingual transfer did not work.