big data and cognitive computing

MDPI

Article

Data Sorting Influence on Short Text Manual Labeling Quality for Hierarchical Classification

Olga Narushynska, Vasyl Teslyuk, Anastasiya Doroshenko * and Maksym Arzubov

Department of Automated Control Systems, Lviv Polytechnic National University, 79013 Lviv, Ukraine; olha.o.narushynska@lpnu.ua (O.N.); vasyl.m.teslyuk@lpnu.ua (V.T.); maksym.v.arzubov@lpnu.ua (M.A.)
* Correspondence: anaastasiia.v.doroshenko@lpnu.ua

Abstract: The precise categorization of brief texts holds significant importance in various applications within the ever-changing realm of artificial intelligence (AI) and natural language processing (NLP). Short texts are everywhere in the digital world, from social media updates to customer reviews and feedback. Nevertheless, short texts' limited length and context pose unique challenges for accurate classification. This research article delves into the influence of data sorting methods on the quality of manual labeling in hierarchical classification, with a particular focus on short texts. The study is set against the backdrop of the increasing reliance on manual labeling in AI and NLP, highlighting its significance in the accuracy of hierarchical text classification. Methodologically, the study integrates AI, notably zero-shot learning, with human annotation processes to examine the efficacy of various data-sorting strategies. The results demonstrate how different sorting approaches impact the accuracy and consistency of manual labeling, a critical aspect of creating high-quality datasets for NLP applications. The study's findings reveal a significant time efficiency improvement in terms of labeling, where ordered manual labeling required 760 min per 1000 samples, compared to 800 min for traditional manual labeling, illustrating the practical benefits of optimized data sorting strategies. Comparatively, ordered manual labeling achieved the highest mean accuracy rates across all hierarchical levels, with figures reaching up to 99% for segments, 95% for families, 92% for classes, and 90% for bricks, underscoring the efficiency of structured data sorting. It offers valuable insights and practical guidelines for improving labeling quality in hierarchical classification tasks, thereby advancing the precision of text analysis in AI-driven research. This abstract encapsulates the article's background, methods, results, and conclusions, providing a comprehensive yet succinct study overview.

Keywords: multiclass hierarchical classification; inter annotator agreement (IAA); automated data labeling; manual data labeling; hierarchical classification; zero-shot learning; short text classification; data annotation techniques; labeling process optimization; hierarchical data organization

Citation: Narushynska, O.; Teslyuk, V.; Doroshenko, A.; Arzubov, M. Data Sorting Influence on Short Text Manual Labeling Quality for Hierarchical Classification. *Big Data Cogn. Comput.* **2024**, *8*, 41. https://doi.org/10.3390/bdcc8040041

Academic Editors: Federica Rollo and John Blake

Received: 15 March 2024
Revised: 28 March 2024
Accepted: 3 April 2024
Published: 7 April 2024

1. Introduction

In the rapidly evolving field of AI and NLP, the accurate classification of short texts plays a pivotal role in various applications [1,2]. Among the myriad of techniques used for data annotation, manual labeling stands out as a fundamental approach for enhancing the performance of AI models [3,4]. While extensive research has been conducted on manual labeling techniques and their impact on text classification, a critical gap remains in understanding the influence of data sorting methods on the quality of manual labeling for hierarchical classification tasks [5,6].

While manual labeling is typically associated with extensive texts or well-structured data, the significance of short-text manual labeling is often overlooked [7]. Short texts are ubiquitous in today's digital landscape, ranging from social media posts to product reviews and customer feedback [8]. However, short texts' limited length and context pose unique challenges for accurate classification. Investigating the influence of data sorting on manual

labeling quality, specifically for short texts, is essential for uncovering strategies that can maximize the performance of hierarchical classification models on this data type [9].

Manual labeling is critical in training AI models, especially when automated techniques are insufficient. Understanding how data sorting strategies impact the quality of manual labeling is essential in order to optimize the labeling process and ensure that the labeled data captures the actual underlying characteristics of the short texts [10]. This knowledge can lead to improved training data, subsequently enhancing the performance of hierarchical classification models [11].

Short texts are prone to data sparsity, causing building robust and generalized AI models to be challenging [12]. Manual labeling enables annotators to provide additional information, such as entity linking, topic categorization, and sentiment annotations, which enrich the training data and augment the model's capacity to handle unseen instances and improve generalization [13].

The existing scientific literature has extensively explored the challenges and methodologies surrounding manual labeling, particularly in the context of traditional text classification [14]. Numerous studies have examined crowdsourcing, active learning, and annotation guidelines to improve annotation quality and efficiency [15]. However, the specific area of data sorting techniques and their influence on short text manual labeling quality for hierarchical classification tasks has received limited attention, leaving a crucial research gap that must be addressed [16].

In light of the abovementioned gap, this research aims to investigate the influence of data sorting methods on the quality of short text manual labeling for hierarchical classification [17]. The research task is to evaluate the data sorting technique in order to determine the impact on the accuracy and consistency of manual labeling. The object of this study is the process of labeling short textual data used for hierarchical classification, and the subject is manual labeling methods and approaches [18].

The primary purpose of this work is to shed light on the importance of data sorting techniques in the manual labeling process for hierarchical classification tasks. By addressing this research gap, we aim to enhance the understanding of how sorting methods can contribute to more effective and reliable manual labeling in short text classification scenarios [19]. The main tasks of this research include analyzing the influence of data sorting on annotation quality, investigating the impact of sorting methods on the inter-annotator agreement, and proposing guidelines for the optimization of the data sorting process in hierarchical text classification.

A detailed description of relevant research will be provided in Section 2. In Section 3, we described types of data labeling and compared different hierarchical classification types (Multiclass Classification, Multilabel Classification, and Multi-tag Classification). Additionally, we derive the Multiclass Hierarchical Classification as the preferable type of classification for this task. Also, in Section 3, we have a detailed description of the dataset and metrics used in this research. Therefore, the mean accuracy and inner-annotator agreement were chosen as the most indicative metrics for this task. The methodology of research was described in Section 4. The different methods used in the study are zero-shot learning for automated data labeling, ordering data before manual labeling, and the manual labeling process. Subsequently, in Section 5, we present the results of four different experiments, including the mean accuracy and inner-annotator agreement. Finally, concluding remarks are presented in Section 6.

2. Relevant Research

Recent research by Gilardi et al. [20] sheds light on the challenges and considerations associated with ChatGPT's reliability and consistency in such tasks. To ensure accuracy and reliability, it is recommended that ChatGPT's outputs are validated by comparing them with human-annotated reference data [21]. Validation is crucial in measuring what researchers intend to measure and instilling confidence in the results.

Several issues arise when using ChatGPT for data labeling. Inconsistency in classification output is a notable concern, as even minor changes in prompts or repeating the same input can yield varying results. The randomness of ChatGPT's outputs, influenced by temperature settings and its black-box nature, further hampers the reliability of its classification and annotation results [22]. Thorough validation, including comparisons against human-annotated reference data, is imperative in addressing these concerns and ensuring the accuracy and reliability of ChatGPT's outputs for data labeling. Therefore, crafting clear and precise instructions for data labeling tasks is paramount [23].

The study concludes that the unsupervised use of ChatGPT for data labeling should be avoided. It highlights the necessity of validation and caution to ensure the reliability and validity of ChatGPT's outputs. Considering the issues of inconsistency, randomness, the need for validation, the sensitivity to instruction variations, and the recommendation against unsupervised use, it becomes evident that ChatGPT for data labeling tasks poses significant challenges and limitations [24].

The article on data cascades in high-stakes AI by [24] sheds light on the importance of accurate data annotation and the consequences of poor data documentation. The study highlights instances where inadequate cross-organizational documentation and the neglect of data documentation led to the costly discarding of datasets or the need for data re-collection [25]. The lack of metadata and changes in the schema without understanding the context resulted in the loss of valuable medical robotic data collection. These examples underscore the significance of precise and comprehensive data annotation in preventing data cascades.

The research also emphasizes the need for customizable dataset collection and labeling tools in order to enhance data quality. Such tools can ensure consistent and accurate data annotation, mitigating the risk of data cascades. However, the study identifies upstream challenges in defining dataset requirements, and downstream challenges in monitoring incoming live data and measuring impacts, which often receive insufficient attention. Therefore, there is a pressing need for improved tools and standards to address data annotation challenges and enhance data quality in developing AI systems.

In [26], several important aspects of data labeling are discussed. The debate surrounding expert annotators versus non-expert annotators is highlighted, with arguments for the effectiveness of both approaches. Quality assurance techniques, including gold standard test questions and various models, are proposed to address low-quality contributions in crowdsourcing. Consistency is identified as a reliable proxy variable in the annotation process, acknowledging that the system may exhibit inconsistency due to varying difficulty levels and the non-uniform performance of annotation workers. The article explores different approaches to annotating, such as one-grader, double-graded with conflict resolution, N-graded, and dynamic automatic conflict resolution, each with its own advantages and limitations. The impact of ambiguity in annotation tasks is discussed, emphasizing the potential loss of information and the increased measurement error if ambiguity is ignored [27]. The article also presents a method for estimating the latent inaccuracy of crowdsourcing annotation processes based on the proportion of requests receiving identical labels. Finally, the importance of assumptions for accurate labeling, including a singular ground truth label, the absence of systematic confusion, and memoryless behavioral stability, is emphasized. Overall, the article underscores the need to consider annotator expertise, quality assurance techniques, consistency, and handling ambiguity in order to ensure accurate and reliable data annotations.

The article [28] discusses the measurement of annotator agreement in text classification tasks. The main concept for text classification inner annotator agreement is to measure the level of agreement or disagreement among multiple annotators when assigning labels or categories to text documents. This agreement is important for assessing the reliability and consistency of the annotations, and can provide insights into the difficulty of the classification task and the quality of the annotation guidelines. Various metrics, including Krippendorff's alpha and Cohen's kappa, are utilized to measure the agreement.

Overall, this article contributes to the understanding and implementing of measuring annotator agreement in complex structured, multi-object, and free-text annotation tasks.

The article [29] highlights the importance of considering disagreements in data labeling for relevant research. Disagreements among annotators can significantly impact the quality and reliability of the labeled data. When annotators have conflicting judgments, it introduces ambiguity and uncertainty into the labeling process, leading to inconsistent and unreliable labels.

One consequence of disagreements is difficulty in establishing a "correct" label for a given instance, as reaching a consensus becomes challenging. This lack of consensus can result in inconsistency in model training and evaluation, as different annotators may have varying interpretations or criteria for labeling.

Disagreements also introduce noise into the labeled data, making distinguishing between genuine linguistic phenomena and annotation errors or inconsistencies harder. This noise negatively affects model performance, as they may learn from incorrect or inconsistent labels.

Furthermore, disagreements shed light on the subjective nature of certain NLP tasks, such as sentiment analysis or hate speech detection. These tasks often involve subjective judgments influenced by individual biases and perspectives. Disagreements among annotators reflect the diversity of opinions and interpretations, thus emphasizing the need for robust and inclusive models that can handle different perspectives.

Considering and addressing disagreements in data labeling is crucial to improve the reliability and quality of labeled data. Understanding the impact of disagreements can enhance the performance and generalizability of NLP models in various applications.

3. Materials and Methods

3.1. Types of Data Labeling

The distinction between labeled and unlabeled data becomes important when considering the machine learning model that is being used. Text classification, for example, involves assigning labels to blocks of text based on the sentiment or topic. Data labeling can be performed manually or automatically, with manual labeling involving inspecting and tagging each piece of data via the use of data labeling software.

There are two main approaches to data labeling: automated data labeling and manual data labeling. Manual data labeling allows for close control over the labeled data and the ability to refine the labeling taxonomy. However, it can be highly time-consuming and prone to human error.

Data labeling experts also use AI models for auto-labeling, where a human verifies and corrects the labels generated by the AI. The corrected data are then used to re-train the auto-label AI, which is eventually added to the pool of labeled training data.

In machine learning and classification tasks, hierarchical class labels are used to organize and represent complex relationships between different classes or categories. Instead of having a flat structure where each class is independent, hierarchical class labels introduce a hierarchical or tree-like structure that captures the hierarchical relationships between classes.

3.2. Hierarchical Classification Types

Hierarchical class labels can be applied to different types of classification problems, such as multiclass, multilabel, and multi-tag classification.

3.2.1. Multiclass Classification

In multiclass classification, each instance or example is assigned to one and only one class from a set of mutually exclusive classes. Hierarchical class labels in multiclass classification can be organized into a hierarchy, where each class is a parent or child of other classes [4]. For example, in a classification problem related to animals, we could have a hierarchy with classes like "Animal" as the parent class, and subclasses like "Mammal", "Bird", and "Reptile".

3.2.2. Multilabel Classification

Multilabel classification deals with instances that can belong to multiple classes simultaneously. Hierarchical class labels in multilabel classification can represent multiple levels of class relationships [30]. An example of this in a hierarchical context could be classifying academic papers into different research areas. The hierarchy could be structured with "Science" at the top level, followed by "Physics", "Chemistry", "Biology", etc. at the next level, and then specific subfields like "Quantum Physics", "Organic Chemistry", "Molecular Biology", etc. at the bottom level. Each paper could be associated with multiple research areas, and the hierarchical structure allows these associations to be captured across different levels of granularity.

3.2.3. Multi-Tag Classification

Multi-tag classification is similar to multilabel classification, but this method allows multiple tags to be assigned to each instance, rather than considering them as classes [31]. Hierarchical class labels in multi-tag classification can capture the complex relationships between different tags [32]. An example of this in a hierarchical context could be tagging posts on a social media platform. The hierarchy could be structured with broad categories like "Sports", "Entertainment", "Technology", etc. at the top level, followed by more specific categories like "Football", "Movies", "Smartphones", etc. at the next level. Each post could be associated with multiple tags, and the hierarchical structure allows for these associations to be captured across different levels of specificity [33].

In all these types of classification, the hierarchical structure of class labels provides a way in which we can model and utilize the relationships and dependencies between different classes, thus enhancing the interpretability and flexibility of the classification model. In this article, we discuss Multiclass Hierarchical Classification.

3.3. Dataset

For hierarchical classes, we use the Global Product Classification (GPC) [34]. It is a standardized system developed by GS1 https://www.gs1.org/ (accessed on 15 January 2024), a global standards organization, in order to classify products and services for efficient communication and data exchange in supply chains. The GPC provides a hierarchical structure, categorizing products into segments, families, classes, bricks, and attributes/attribute values (Table 1).

Table 1. Dataset example.

Brick Name	Product Description
Air Fresheners/Deodorizers (Non Powered)	Air Wick Aqua Mist Tropical Bliss Air Refresher
Air Fresheners/Deodorizers (Non Powered)	Air Wick Fresh Matic Spring Blossoms Automatic Ultra Spray Refill
Air Fresheners/Deodorizers (Non Powered)	airBoss Air Freshener Rainfresh
Air Fresheners/Deodorizers (Non Powered)	airBoss Closet Odor Eliminator
Air Fresheners/Deodorizers (Non Powered)	Aura Cacia Aromatherapy Diffuser Refill Pads
Alcoholic Pre-mixed Drinks	Mike's Harder Black Cherry Lemonade
Alcoholic Pre-mixed Drinks	Mike's Harder Black Cherry Lemonade—4 CT
Alcoholic Pre-mixed Drinks	Mr & Mrs T Non-Alcoholic Lemon Tom Collins Mix
Alcoholic Pre-mixed Drinks	Mr. & Mrs. T Non-Alcoholic Pina Colada Mix
Alcoholic Pre-mixed Drinks	Mr. & Mrs. T Non-Alcoholic Strawberry Daiquiri-Margarita Mix
Apple/Pear Alcoholic Beverage—Sparkling	Crispin Honey Crisp Hard Cider
Apple/Pear Alcoholic Beverage—Sparkling	Crispin Honey Crisp Hard Cider
Apple/Pear Alcoholic Beverage—Sparkling	Hornsby's Hard Cider Amber Draft—6 PK
Bakeware/Ovenware/Grillware (Non Disposable)	Pyrex Vintage Charm Bowl Set Birds Of A Feather—3 CT
Bakeware/Ovenware/Grillware (Non Disposable)	Pyrex Vintage Charm Bowl Set Rise N' Shine—3 CT
Bakeware/Ovenware/Grillware (Non Disposable)	Rachael Ray 2 Piece Set Ceramic Bakeware Oval Bakers 1 ea BOX
Bakeware/Ovenware/Grillware (Non Disposable)	Rachael Ray Baker 1 ea BOX
Bakeware/Ovenware/Grillware (Non Disposable)	Rachael Ray Bakerware 1 ea CARD
Bakeware/Ovenware/Grillware (Non Disposable)	Rachael Ray Bakeware 1 ea BOX

An example of a taxonomy for milk can be found in the GPC system. In the GPC, milk falls under the segment "Food and Beverage Products" (Segment ID: 10000000). Within this segment, milk is classified under the family "Dairy Products" (Family ID: 10000001). The specific class for milk may vary depending on the GPC version and language, but it is typically categorized under the class "Liquid milk" (Class ID: 10000002).

In our study, we utilize a unique dataset derived from web scraping activities on the DirectionsForMe website [35]. Each data entry in this dataset comprises a textual product description and a corresponding category label, termed a 'brick', from the GPC [36]. Our dataset provides a rich source of information for analyzing product descriptions and their associated GPC categories, offering valuable insights into product classification and consumer behavior patterns.

In order to rigorously evaluate the performance of different methods, we partitioned our dataset into three distinct groups (dataset-1, dataset-2, dataset-3), each containing an equal number of samples per brick (10 samples). These groups spanned 3 segments (Beauty/Personal Care/Hygiene, Food/Beverage/Tobacco, and Kitchenware and Tableware), 14 families, 32 classes, and 100 bricks, resulting in 1000 samples per group. This stratified sampling approach ensures a balanced representation of each category across all three groups, thereby enabling a fair and robust comparison of the methods being tested [37].

Using ground truth labels in our dataset is paramount for our research. Ground truth labels, in this case, the GPC 'brick' labels, serve as the definitive, known classifications for each product description. These labels provide a reliable standard against which the annotation of our tested methods can be compared and evaluated. By having a ground truth, we can objectively measure the accuracy of each method in labeling product descriptions, identify their strengths and weaknesses, and guide the refinement of these methods.

3.4. Metrics

Measuring the accuracy of our labeling approaches at each level of the GPC (i.e., segment, family, class, and brick) is crucial for a comprehensive understanding of the performance of our methods. Accuracy (1) is an especially suitable metric for our study due to the nature of our dataset and our research objectives [38]. As our dataset is balanced, with an equal number of samples per brick, accuracy avoids the potential pitfalls associated with imbalanced datasets, where a high accuracy can be achieved by simply predicting the majority class. Furthermore, our goal is to evaluate the overall correctness of our labeling methods across all categories, rather than focusing on the performance within specific categories or the balance between precision and recall. In this context, accuracy provides a straightforward and intuitive measure of performance, reflecting the proportion of product descriptions that are correctly classified. While other metrics, such as the F1 score, might be more suitable in cases where false positives and false negatives have different costs or where the dataset is imbalanced, the simplicity and interpretability of accuracy make it an ideal choice for our study.

$$\text{Accuracy}_{\text{annotator}} = \frac{\text{correct labels}}{\text{all samples}}, \tag{1}$$

Each level of the GPC hierarchy provides a different granularity of product categorization, and the accuracy at each level can shed light on how well our methods can handle different levels of specificity. For instance, high accuracy at the 'segment' level but lower accuracy at the 'brick' level may indicate that our methods are better at broad categorization, but need help with fine-grained classification [39]. We can pinpoint where improvements are needed by evaluating the accuracy at each level, tailoring our methods accordingly.

Calculating the mean accuracy (2) across all annotators provides a holistic view of the performance of our labeling methods. This metric averages the accuracy of each annotator, giving us a single measure that encapsulates the overall effectiveness of our methods at

different levels of product categorization. It allows us to identify which levels our annotators are most accurate at and where they struggle, guiding us in refining our labeling process.

$$\text{Mean Accuracy} = \frac{1}{n}\sum_{n}^{1} \text{Accuracy}_{\text{n annotator}} \quad (2)$$

On the other hand, the inter-annotator agreement measures the consistency between different annotators or labeling methods. In our study, it serves as an indicator of the reliability and reproducibility of our labeling process. A high inter-annotator agreement suggests that our labeling methods are robust and yield consistent results, regardless of the specific method or annotator used.

Fleiss' Kappa (k) is a statistical measure used to assess the reliability of agreement between multiple raters when assigning categorical ratings to a number of items or classifying items (3). It is an extension of Cohen's Kappa, which is used for the same purpose, but only works for two raters. Fleiss' Kappa, on the other hand, works for any fixed number of raters.

$$k = \frac{p_a - p_e}{1 - p_e} \quad (3)$$

where p_a—the observed proportion of agreement, and p_e—is the expected proportion due to chance agreement.

Like most correlation coefficients, k ranges from 0 to 1, where

- 0 is no agreement (or agreement that you would expect to find by chance),
- 1 is a perfect agreement.

This is particularly important in research, as it provides our findings with credibility, ensuring that they can be reliably replicated in future studies. Furthermore, by comparing inter-annotator agreement and accuracy, we can gain insights into the trade-off between consistency (agreement) and correctness (accuracy), guiding us in refining our labeling approaches.

4. Methodology

4.1. Zero-Shot Learning for Automated Data Labeling

Zero-shot learning presents a promising approach to the challenge of automatic labeling in short text datasets [40]. This technique, which allows a machine learning model to classify data accurately that it has never seen before, is particularly well suited to short text data's dynamic and diverse nature [41,42]. One of the most powerful tools for zero-shot learning in NLP is the Bidirectional and Auto-Regressive Transformers (BART) model from Hugging Face, specifically the 'bart-large-mnli' variant [43].

The BART model is a denoising autoencoder for pretraining sequence-to-sequence models [44]. It works by corrupting the input text and then learning to reconstruct the original data [45]. The 'bart-large-mnli' model is trained on the Multi-Genre Natural Language Inference (MNLI) dataset, a large-scale, crowd-sourced collection of sentence pairs annotated with textual entailment information [46].

In the context of short text dataset labeling, the BART model can be used to predict the label of a given text based on its semantic similarity to predefined label descriptions. The model treats the task as a textual entailment problem, where the goal is to determine whether a given label description 'entails' the meaning of the text. This approach allows the model to assign labels to texts based on their inherent semantic content without requiring any labeled examples of each class in the training data [47].

In conclusion, the 'bart-large-mnli' model from Hugging Face offers a powerful and flexible tool for automatically labeling short text datasets. Its ability to perform zero-shot learning enables it to easily adapt to new tasks and labels, making it a valuable asset in the rapidly evolving field of NLP [48].

Employing the 'bart-large-mnli' model for zero-shot learning in short text dataset labeling removes the need for inter-annotator agreement measures [44]. The labeling, performed by a single deterministic model, ensures consistency, simplifying the evaluation process [49].

The primary metric becomes accuracy, providing a straightforward measure of the model's performance. However, to ensure the highest level of accuracy, it is beneficial to incorporate a 'human in the loop' approach. This allows for the manual reviewing and correcting of the model's predictions, further enhancing the quality and reliability of the labeling.

4.2. Ordering Data before Manual Labeling

In our study, we explored the impact of leaf ordering in hierarchical clustering on the accuracy of our labeling process. Leaf ordering is a technique used in this context to rearrange the leaves in the resulting dendrogram to reveal the underlying structure of the data more clearly [50].

Leaf ordering can significantly impact labeling accuracy in text data. By rearranging the leaves, we can group similar texts more closely together [51]. This not only orders texts with similar meanings (and thus similar 'bricks') close to each other, but also helps assign more accurate labels, especially when the labels are hierarchical. This is particularly relevant in our case, where we deal with hierarchical class labels from the GPC system.

We used BART for text embedding to facilitate the clustering and labeling process. These representations capture the semantic content of the text, allowing us to measure the similarity between different texts and group them accordingly.

Combining BART embeddings with leaf ordering in hierarchical clustering enhanced the accuracy of our labeling process and maximized inner-annotator agreement. Therefore, this combination of techniques provided a powerful tool for the manual labeling of our short text dataset.

The introduction of leaf ordering in hierarchical clustering within our study is primarily motivated by its ability to improve data interpretability for manual labeling. Arranging similar items closely in the cluster tree aids labelers in recognizing patterns more efficiently, thereby enhancing the accuracy of manual labeling. We selected this approach for its adaptability across various hierarchical classification tasks, as we appreciated its potential to streamline the labeling process in diverse contexts. It is important to note that selecting the best ordering algorithm was not the aim of our article; instead, we focused on showcasing how data ordering can benefit manual labeling in hierarchical classification.

4.3. Manual Labeling Process

The labeling tool we use is Label Studio, with a custom XML configuration reflecting our taxonomy.

```
<View>
<Text name="text" value="$text"/>
<Taxonomy name="taxonomy" showFullPath="true" leafsOnly="true" maxUsages="1"
toName="text">
<Choice value="FoodBeverageTobacco">
<Choice value="Food">
<Choice value="Bakery">
<Choice value="Bread"/>
<Choice value="Pastries"/>
<Choice value="Cakes"/>
</Choice>
<Choice value="Dairy">
<Choice value="Milk"/>
<Choice value="Cheese"/>
<Choice value="Yogurt"/>
</Choice>
<Choice value="FruitsVegetables">
<Choice value="Fruits"/>
<Choice value="Vegetables"/>
<Choice value="Herbs"/>
```

</Choice>
</Choice>
</Choice>
Instructions for Data Labeling

1. Set up Label Studio with the XML configuration adjusted to match the GPC standard taxonomy (Segment, Family, Class, and Brick).
2. Import the dataset (ordered or not), which consists of text samples that require annotation (Figure 1).
3. When labeling item descriptions based on the GPC standard, following the hierarchical structure of Segments, Families, Classes, and Bricks is essential. Here is a detailed step-by-step guide on how to label item descriptions using the GPC (Figure 2):

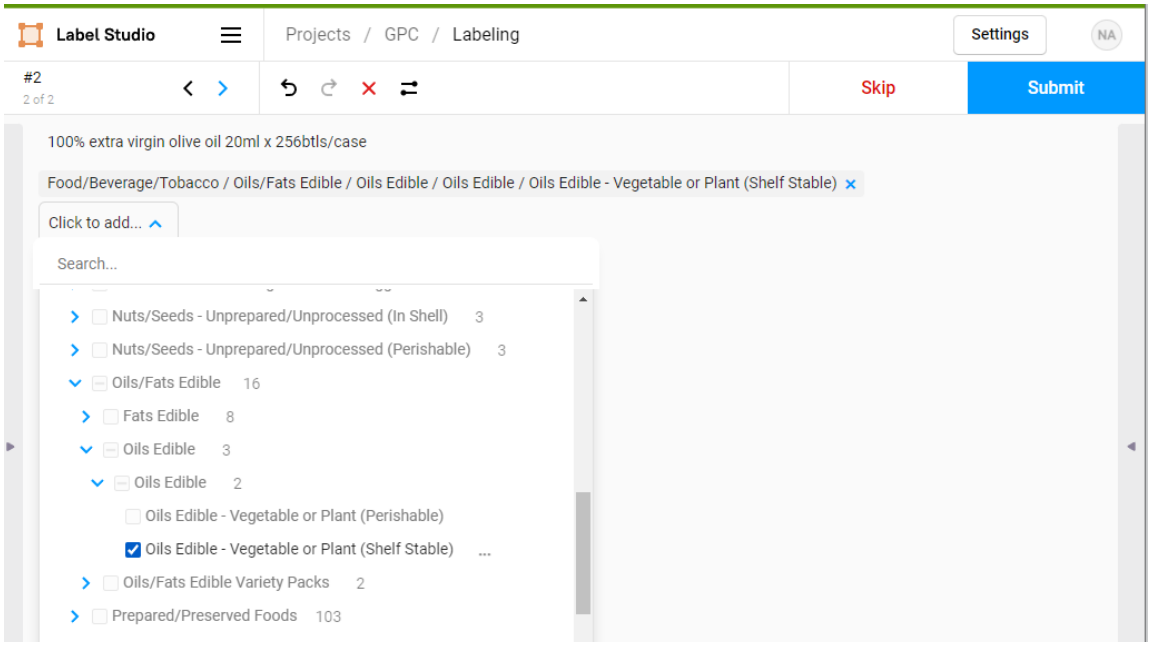

Figure 1. Process of data labeling in Label Studio.

3.1. Understand the item description: Carefully read the item description and note any critical information, such as ingredients, materials, intended use, or other relevant attributes. This information will help you determine the most suitable GPC categories.
3.2. Identify the Segment: Start by selecting the Segment (top-level category) that best describes the item. Segments are broad categories that cover a wide range of products. Examples of Segments include Food and Beverage, Cleaning Products, and Self Care. Choose the Segment that is most relevant to the item description.
3.3. Select the Family: Within the Segment, identify the Family (subcategory) to which the item belongs. Families are more specific than Segments and provide a narrower scope of categorization. For example, within the Food and Beverage Segment, you may have Families like Dairy, Bakery, and Processed Meat Products. Choose the Family that best matches the item description.
3.4. Determine the Class: Identify the Class (sub-subcategory) under which the item falls within the selected Family. Classes further subdivide Families, providing a more refined categorization. For example, you may have Cheese, Milk, and Yogurt Classes within the Dairy Products Family. Select the Class that corresponds most closely to the item description.
3.5. Choose the Brick: Finally, within the chosen Class, select the Brick (finest level of categorization) that best describes the item. Bricks represent the most specific and granular level of the GPC taxonomy. For instance, within the Cheese Class, you may

have Bricks like Blue Cheese, Cheddar Cheese, and Mozzarella Cheese. Choose the Brick that accurately represents the item based on the description.

3.6. Maintain consistency: Ensure you follow the GPC structure and guidelines consistently throughout the labeling process. Each item should be labeled with one Brick. In cases where multiple Bricks may seem applicable, choose the one that best represents the item's primary characteristics or intended use.

3.7. Quality control: Regularly review your labeling progress and check for any inconsistencies or errors. If necessary, correct any mistakes and ensure that the labeled dataset is accurate and consistent with the GPC standard.

3.8. When encountering unknown items or items that do not seem to fit into the existing GPC taxonomy, follow these steps to handle them:

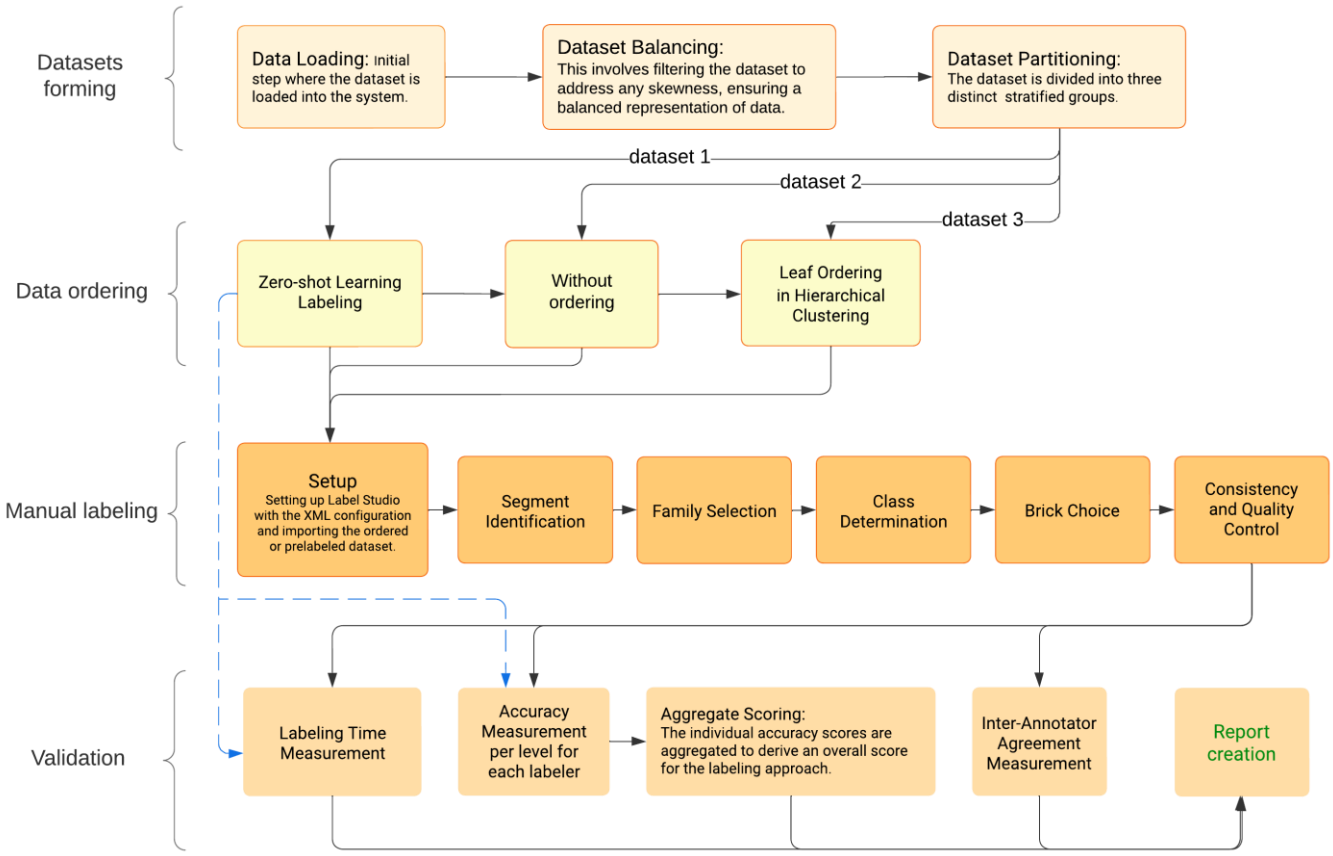

Figure 2. Detailed workflow of the data labeling process.

- Re-examine the item description: Carefully reread the item description and ensure you have not missed any crucial information that could help identify the appropriate category. Sometimes, important details might be overlooked in the initial review.
- Research the item: If the item is unfamiliar, conduct some research in order to gain a better understanding of the product. This may include searching the internet, referring to product catalogs, or consulting with experts. The additional information gathered can help determine the most suitable category.
- Flag the item: If you are unable to determine an appropriate category even after researching and considering related categories, flag the item for further review. This can be completed using a specific label, such as "Unknown" or "Requires Review". These flagged items can then be revisited and discussed with your team, supervisor, or subject matter expert to determine the best course of action.

5. Results and Discussion

In order to compare the accuracy of the proposed approaches to data labeling, a series of experiments was conducted.

5.1. Experiment 1: Zero-Shot Learning

We applied zero-shot learning to label dataset-1 in the first experiment automatically. We utilized the 'bart-large-mnli' model from Hugging Face, which is capable of classifying texts into categories based on their inherent semantic content without requiring any labeled examples of each class in the training data. This experiment served as a baseline, providing an initial measure of the feasibility and effectiveness of automatic labeling for our dataset. Zero-shot learning, central to our methodology, allows for the dynamic classification of text based on conceptual similarity and context rather than relying on a pre-defined label set, thus significantly enhancing the flexibility and scalability of the manual labeling processes in hierarchical classification tasks.

The hierarchical application of zero-shot learning, from the broadest categories at the Segments level down to the more specific entities at the Bricks level, enables a nuanced and layered approach to classification. By applying zero-shot learning in this top-down manner, we ensure that each subsequent level of classification is informed by the context and categorizations established at the higher levels, thereby optimizing the coherence and accuracy of the labeling process across the hierarchy.

5.2. Experiment 2: Manual Labeling

We employed a manual labeling process for dataset-2 in the second experiment. Five annotators independently assigned labels to each text in the dataset. This experiment allowed us to assess the reliability and consistency of manual labeling, as measured by the inter-annotator agreement, and to compare its accuracy with that of the automatic labeling from the first experiment.

5.3. Experiment 3: Manual Correction

We introduced a 'human in the loop' approach in the third experiment. The same five annotators from the second experiment manually reviewed and corrected the labels that dataset-1 assigned via the automatic method used in the first experiment. This experiment aimed to combine the strengths of automatic and manual labeling, leveraging the former's speed and consistency with the latter's nuance and expertise.

In our combined approach of manual checks on zero-shot automatic labeling, the decision to incorporate manual intervention is guided by the principle of using the zero-shot model's output as a suggestion rather than a definitive classification. This approach acknowledges the inherent limitations of even the most advanced AI models in capturing the full context and nuances of hierarchical classification tasks, especially in domains with highly specialized or nuanced categories.

This flexible combined methodology allows us to leverage the efficiency and scalability of zero-shot learning while maintaining the accuracy and nuance afforded by human expertise. It offers a practical framework for researchers applying similar methods, balancing the strengths of automated and manual labeling processes to achieve high-quality data classification.

5.4. Experiment 4: Manual Labeling with Ordered Data

In the fourth and final experiment, we again employed manual labeling by the same five annotators, but this time on dataset-3, which had been ordered using leaf ordering in hierarchical clustering. This experiment aimed to assess the impact of data ordering on the labeling process. By grouping similar texts, we hypothesized that the annotators would be able to assign labels more accurately and consistently.

Through these four experiments, we aimed to explore different approaches to labeling, assess their strengths and weaknesses, and identify the most effective methods for our specific task.

The results of our experiments highlight the trade-offs between different labeling approaches in terms of the labeling time per 1000 samples (Table 2).

Table 2. Time trade-offs between different labeling approaches.

Labeling Approach	Labeling Time (per 1000 Samples), mins
Zero-shot automatic labeling	37
Manual check zero-shot automatic labeling	600
Manual labeling	800
Ordered manual labeling	760

The fastest approach was the zero-shot automatic labeling, which took only 37 min to label 1000 samples. This underscores the efficiency of automatic labeling, which can process large volumes of data quickly due to the lack of manual intervention.

However, when a manual check was introduced to the zero-shot automatic labeling in the third experiment, the labeling time increased significantly to 600 min. This increase reflects the additional time required for human annotators to review and correct the labels assigned by the automatic method. Despite this increase, the combined automatic–manual approach was still faster than pure manual labeling.

In the second experiment, pure manual labeling took the longest time, at 800 min for 1000 samples. This result highlights the time-consuming nature of manual labeling, which requires human annotators to read and understand each text and assign an appropriate label.

In the fourth experiment, the manual labeling of ordered data was slightly faster than pure manual labeling, taking 760 min for 1000 samples. This suggests that ordering the data can improve the efficiency of manual labeling, likely by grouping similar texts together, thereby reducing the cognitive load on the annotators.

4. The zero-shot automatic labeling approach was the least accurate, but was also the fastest. Manual labeling significantly improved the accuracy across all GPC hierarchy levels (Table 3). The manual checking of zero-shot automatic labeling further increased the accuracy, suggesting that combining automatic and manual methods can balance speed and accuracy (Figure 3).

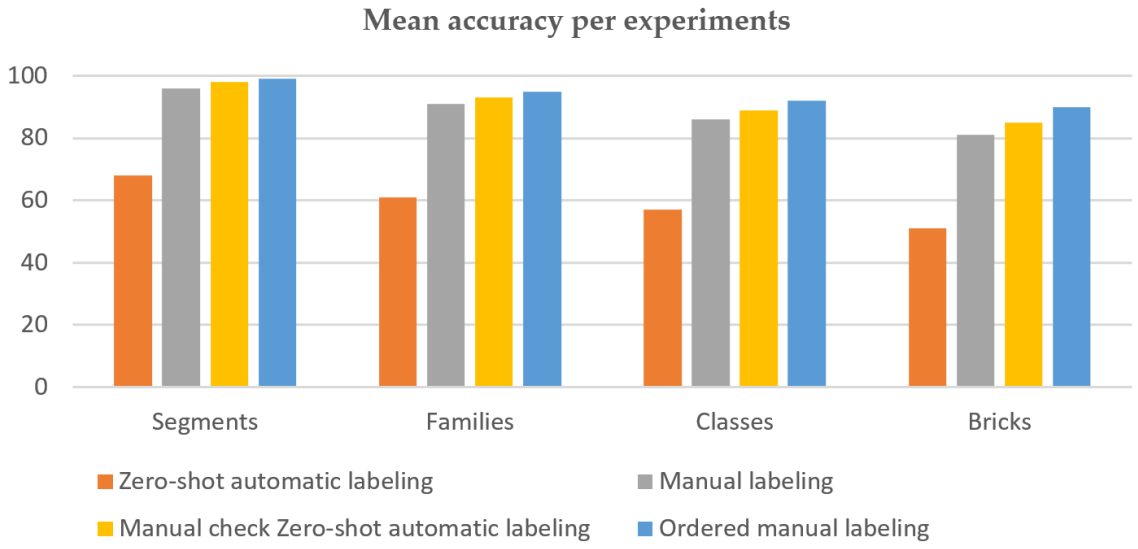

Figure 3. Mean accuracy data chart of different approaches.

Finally, ordered manual labeling achieved the highest accuracy, indicating that data ordering can enhance manual labeling efficiency and accuracy. Thus, the optimal labeling method may depend on the specific balance between speed and accuracy required for the task.

Table 3. Mean accuracy per experiment.

Approach	Mean Accuracy (%)			
	Segments	**Families**	**Classes**	**Bricks**
Zero-shot automatic labeling	68	61	57	51
Manual labeling	96	91	86	81
Manual check zero-shot automatic labeling	98	93	89	85
Ordered manual labeling	99	95	92	90

The inter-annotator agreement results provide insight into the consistency of the labeling process across different annotators (Table 4).

Table 4. Inner-annotator agreement per experiments.

Approach	Inner-Annotator Agreement
Manual labeling	0.82
Manual check zero-shot automatic labeling	0.87
Ordered manual labeling	0.91

Manual labeling resulted in an agreement score of 0.82, indicating a high level of consistency among the annotators. The manual check of zero-shot automatic labeling further improved the agreement to 0.87, suggesting that combining automatic and manual labeling can enhance consistency. Finally, ordered manual labeling achieved the highest agreement score of 0.91, indicating that ordering the data can not only enhance accuracy, as seen in the previous results, but can also improve the consistency of manual labeling.

6. Conclusions

This research explored various approaches to labeling short text datasets, specifically focusing on using zero-shot learning, manual labeling, and a combination of both. The results highlighted the trade-offs between speed, accuracy, and consistency in the labeling process.

Zero-shot automatic labeling, while the fastest method, achieved the lowest accuracy and did not require inter-annotator agreement measures. Manual labeling, although time-consuming, significantly improved the accuracy and demonstrated high inter-annotator agreement, thus emphasizing the value of human expertise in the labeling process.

Combining automatic and manual labeling, where manual checks were introduced to zero-shot automatic labeling, provided a balance between speed and accuracy. This approach also improved the inter-annotator agreement, indicating enhanced consistency in the labeling process.

The introduction of leaf ordering in hierarchical clustering for manual labeling further improved both the accuracy and inter-annotator agreement. This suggests that data ordering can enhance the efficiency and consistency of manual labeling, making it a promising strategy for optimizing the labeling process.

In conclusion, the optimal labeling method depends on the specific requirements of the task. While automatic labeling offers speed, manual labeling provides higher accuracy and consistency. A combined approach or the use of data ordering techniques can offer a balance between these factors, enhancing the overall effectiveness of the labeling process. Future research may explore these methods in more depth or investigate other strategies for optimizing the labeling of short-text datasets.

This article underscores the critical role of data sorting, while also acknowledging the inherent limitations and obstacles tied to our proposed methodologies. Variations in the nature of the data, particularly across different textual domains, might impact the efficacy of sorting strategies due to the distinct characteristics of each domain. Furthermore, the embedding methods employed for sorting warrant careful consideration, as their effectiveness can differ based on the specific requirements of the text being analyzed.

Additionally, the need for significant computational resources to execute complex sorting and embedding processes could pose challenges, especially for those with constrained computational access. By discussing these limitations, we aim to provide a more rounded perspective on our methodology, paving the way for future research to address these issues.

To address the practical implications of our research findings for real-world applications, it is essential to emphasize how the methodologies explored in this study can be seamlessly integrated into various operational environments to enhance the efficiency and accuracy of short-text dataset labeling. Our exploration of zero-shot learning, manual labeling, and their combination, alongside the use of leaf ordering in hierarchical clustering, provides a comprehensive toolkit that can be adapted to improve data processing workflows in different industries, ranging from customer service and social media monitoring to content categorization and sentiment analysis.

Author Contributions: Conceptualization, O.N. and A.D.; methodology, O.N.; software, M.A.; validation, O.N., V.T. and M.A.; formal analysis, O.N. and A.D.; investigation, O.N.; resources, V.T.; data curation, O.N. and V.T.; writing—original draft preparation, O.N.; writing—review and editing, A.D.; visualization, A.D.; supervision, V.T.; project administration, M.A.; funding acquisition, V.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data supporting reported results can be found at https://www.directionsforme.org/categories (accessed on 15 January 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kasperiuniene, J.; Briediene, M.; Zydziunaite, V. Automatic Content Analysis of Social Media Short Texts: Scoping Review of Methods and Tools. In *Computer Supported Qualitative Research*; Costa, A.P., Reis, L.P., Moreira, A., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2020; Volume 1068, pp. 89–101, ISBN 978-3-030-31786-7.
2. Maddigan, P.; Susnjak, T. Chat2VIS: Generating Data Visualizations via Natural Language Using ChatGPT, Codex and GPT-3 Large Language Models. *IEEE Access* **2023**, *11*, 45181–45193. [CrossRef]
3. Zhou, X.; Wu, T.; Chen, H.; Yang, Q.; He, X. Automatic Annotation of Text Classification Data Set in Specific Field Using Named Entity Recognition. In Proceedings of the 2019 IEEE 19th International Conference on Communication Technology (ICCT), Xi'an, China, 16–19 October 2019; IEEE: Xi'an, China, 2019; pp. 1403–1407.
4. Doroshenko, A.; Tkachenko, R. Classification of Imbalanced Classes Using the Committee of Neural Networks. In Proceedings of the 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 11–14 September 2018; IEEE: Lviv, Ukraine, 2018; pp. 400–403.
5. Chang, C.-M.; Mishra, S.D.; Igarashi, T. A Hierarchical Task Assignment for Manual Image Labeling. In Proceedings of the 2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), Memphis, TN, USA, 14–18 October 2019; IEEE: Memphis, TN, USA, 2019; pp. 139–143.
6. Savchuk, D.; Doroshenko, A. Investigation of Machine Learning Classification Methods Effectiveness. In Proceedings of the 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 22–25 September 2021; IEEE: Lviv, Ukraine, 2021; pp. 33–37.
7. Bogatinovski, J.; Todorovski, L.; Džeroski, S.; Kocev, D. Comprehensive Comparative Study of Multi-Label Classification Methods. *Expert Syst. Appl.* **2022**, *203*, 117215. [CrossRef]
8. Nava-Muñoz, S.; Graff, M.; Escalante, H.J. Analysis of Systems' Performance in Natural Language Processing Competitions. *arXiv* **2024**, arXiv:2403.04693. [CrossRef]
9. Ferrandin, M.; Cerri, R. Multi-Label Classification via Closed Frequent Labelsets and Label Taxonomies. *Soft Comput.* **2023**, *27*, 8627–8660. [CrossRef]
10. Narushynska, O.; Teslyuk, V.; Vovchuk, B.-D. Search Model of Customer's Optimal Route in the Store Based on Algorithm of Machine Learning A*. In Proceedings of the 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), Lviv, Ukraine, 5–8 September 2017; IEEE: Lviv, Ukraine, 2017; pp. 284–287.
11. Chang, C.; Zhang, J.; Ge, J.; Zhang, Z.; Wei, J.; Li, L. Interaction-Based Driving Scenario Classification and Labeling. *arXiv* **2024**, arXiv:2402.07720. [CrossRef]
12. Xu, X.; Li, B.; Shen, Y.; Luo, B.; Zhang, C.; Hao, F. Short Text Classification Based on Hierarchical Heterogeneous Graph and LDA Fusion. *Electronics* **2023**, *12*, 2560. [CrossRef]

13. Tang, H.; Kamei, S.; Morimoto, Y. Data Augmentation Methods for Enhancing Robustness in Text Classification Tasks. *Algorithms* **2023**, *16*, 59. [CrossRef]
14. Omar, M.; Choi, S.; Nyang, D.; Mohaisen, D. Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions. *IEEE Access* **2022**, *10*, 86038–86056. [CrossRef]
15. Jin, R.; Du, J.; Huang, W.; Liu, W.; Luan, J.; Wang, B.; Xiong, D. A Comprehensive Evaluation of Quantization Strategies for Large Language Models. *arXiv* **2024**, arXiv:2402.16775. [CrossRef]
16. Peng, Z.; Abdollahi, B.; Xie, M.; Fang, Y. Multi-Label Classification of Short Texts with Label Correlated Recurrent Neural Networks. In Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval, Virtual Event, Canada, 11–15 July 2021; ACM: New York, NY, USA, 2021; pp. 119–122.
17. Arzubov, M.; Narushynska, O.; Batyuk, A.; Cherkas, N. Concept of Server-Side Clusterization of Semi-Static Big Geodata for Web Maps. In Proceedings of the 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, 19–21 October 2023; IEEE: Lviv, Ukraine, 2023; pp. 1–4.
18. Chen, M.; Ubul, K.; Xu, X.; Aysa, A.; Muhammat, M. Connecting Text Classification with Image Classification: A New Preprocessing Method for Implicit Sentiment Text Classification. *Sensors* **2022**, *22*, 1899. [CrossRef] [PubMed]
19. Bercaru, G.; Truică, C.-O.; Chiru, C.-G.; Rebedea, T. Improving Intent Classification Using Unlabeled Data from Large Corpora. *Mathematics* **2023**, *11*, 769. [CrossRef]
20. Gilardi, F.; Alizadeh, M.; Kubli, M. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *arXiv* **2023**, arXiv:2303.15056. [CrossRef]
21. Qiu, Y.; Jin, Y. ChatGPT and Finetuned BERT: A Comparative Study for Developing Intelligent Design Support Systems. *Intell. Syst. Appl.* **2024**, *21*, 200308. [CrossRef]
22. Shah, A.; Chava, S. Zero Is Not Hero Yet: Benchmarking Zero-Shot Performance of LLMs for Financial Tasks. *arXiv* **2023**, arXiv:2305.16633. [CrossRef]
23. Reiss, M.V. Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. *arXiv* **2023**, arXiv:2304.11085. [CrossRef]
24. Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.K.; Aroyo, L.M. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021.
25. Troxler, A.; Schelldorfer, J. Actuarial Applications of Natural Language Processing Using Transformers: Case Studies for Using Text Features in an Actuarial Context. *arXiv* **2022**, arXiv:2206.02014. [CrossRef]
26. Sun, D.Q.; Kotek, H.; Klein, C.; Gupta, M.; Li, W.; Williams, J.D. Improving Human-Labeled Data through Dynamic Automatic Conflict Resolution. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; Scott, D., Bel, N., Zong, C., Eds.; International Committee on Computational Linguistics: Barcelona, Spain, 2020; pp. 3547–3557.
27. Akay, H.; Kim, S.-G. Extracting Functional Requirements from Design Documentation Using Machine Learning. *Procedia CIRP* **2021**, *100*, 31–36. [CrossRef]
28. Braylan, A.; Alonso, O.; Lease, M. Measuring Annotator Agreement Generally across Complex Structured, Multi-Object, and Free-Text Annotation Tasks. In Proceedings of the ACM Web Conference 2022, Lyon, France, 25–29 April 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1720–1730.
29. Basile, V.; Fell, M.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; Poesio, M.; Uma, A. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*; Church, K., Liberman, M., Kordoni, V., Eds.; Association for Computational Linguistics: Melbourne, Australia, 2021; pp. 15–21.
30. Zhu, Y.; Zamani, H. ICXML: An In-Context Learning Framework for Zero-Shot Extreme Multi-Label Classification. *arXiv* **2023**, arXiv:2311.09649. [CrossRef]
31. Doshi, I.; Sajjalla, S.; Choudhari, J.; Bhatt, R.; Dasgupta, A. Efficient Hierarchical Clustering for Classification and Anomaly Detection. *arXiv* **2020**, arXiv:2008.10828. [CrossRef]
32. Kasundra, J.; Schulz, C.; Mirsafian, M.; Skylaki, S. A Framework for Monitoring and Retraining Language Models in Real-World Applications. *arXiv* **2023**, arXiv:2311.09930. [CrossRef]
33. Xu, H.; Chen, M.; Huang, L.; Vucetic, S.; Yin, W. X-Shot: A Unified System to Handle Frequent, Few-Shot and Zero-Shot Learning Simultaneously in Classification. *arXiv* **2024**, arXiv:2403.03863. [CrossRef]
34. Global Product Classification (GPC). Available online: https://www.gs1.org/standards/gpc (accessed on 7 December 2023).
35. Directionsforme. Available online: https://www.directionsforme.org/ (accessed on 7 December 2023).
36. Martorana, M.; Kuhn, T.; Stork, L.; van Ossenbruggen, J. Text Classification of Column Headers with a Controlled Vocabulary: Leveraging LLMs for Metadata Enrichment. *arXiv* **2024**, arXiv:2403.00884. [CrossRef]
37. Miranda, L.J.V. Developing a Named Entity Recognition Dataset for Tagalog. *arXiv* **2023**, arXiv:2311.07161. [CrossRef]
38. Lukasik, M.; Narasimhan, H.; Menon, A.K.; Yu, F.; Kumar, S. Metric-Aware LLM Inference. *arXiv* **2024**, arXiv:2403.04182. [CrossRef]

39. Doroshenko, A. Application of global optimization methods to increase the accuracy of classification in the data mining tasks. In *Computer Modeling and Intelligent Systems. Proceedings of the 2nd International Conference CMIS-2019, Vol-2353: Main Conference, Zaporizhzhia, Ukraine, 15–19 April 2019*; Luengo, D., Subbotin, S., Eds.; 2019; pp. 98–109. Available online: http://ceur-ws.org/Vol-2353/ (accessed on 7 December 2023).
40. Alhoshan, W.; Ferrari, A.; Zhao, L. Zero-Shot Learning for Requirements Classification: An Exploratory Study. *Inf. Softw. Technol.* **2023**, *159*, 107202. [CrossRef]
41. Rondinelli, A.; Bongiovanni, L.; Basile, V. Zero-Shot Topic Labeling for Hazard Classification. *Information* **2022**, *13*, 444. [CrossRef]
42. Zhang, Y.; Wang, Y.; Zhang, H.; Zhu, B.; Chen, S.; Zhang, D. OneLabeler: A Flexible System for Building Data Labeling Tools. In Proceedings of the CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 27 April 2022; ACM: New Orleans, LA, USA, 2022; pp. 1–22.
43. Zhao, X.; Ouyang, S.; Yu, Z.; Wu, M.; Li, L. Pre-Trained Language Models Can Be Fully Zero-Shot Learners. *arXiv* **2022**, arXiv:2212.06950. [CrossRef]
44. Yadav, S.; Kaushik, A.; McDaid, K. Leveraging Weakly Annotated Data for Hate Speech Detection in Code-Mixed Hinglish: A Feasibility-Driven Transfer Learning Approach with Large Language Models. *arXiv* **2024**, arXiv:2403.02121. [CrossRef]
45. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
46. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension. *arXiv* **2019**, arXiv:1910.13461.
47. Meng, Y.; Huang, J.; Zhang, Y.; Han, J. Generating Training Data with Language Models: Towards Zero-Shot Language Understanding. In *Proceedings of the Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 462–477.
48. Srivastava, S.; Labutov, I.; Mitchell, T. Zero-Shot Learning of Classifiers from Natural Language Quantification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Gurevych, I., Miyao, Y., Eds.; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 306–316.
49. Vidra, N.; Clifford, T.; Jijo, K.; Chung, E.; Zhang, L. Improving Classification Performance With Human Feedback: Label a Few, We Label the Rest. *arXiv* **2024**, arXiv:2401.09555. [CrossRef]
50. Bar-Joseph, Z.; Gifford, D.K.; Jaakkola, T.S. Fast Optimal Leaf Ordering for Hierarchical Clustering. *Bioinformatics* **2001**, *17*, S22–S29. [CrossRef] [PubMed]
51. Novoselova, N.; Wang, J.; Klawonn, F. Optimized Leaf Ordering with Class Labels for Hierarchical Clustering. *J. Bioinform. Comput. Biol.* **2015**, *13*, 1550012. [CrossRef]