2024

Age Ain't Just a Number: Exploring the Volume vs. Age Dilemma for Textual Data to Enhance Decision Making

Lukas Hägele
University of Ulm, Germany, lukas.haegele@uni-ulm.de

Mathias Klier
University of Ulm, Germany, mathias.klier@uni-ulm.de

Andreas Obermeier
University of Ulm, Germany, andreas.obermeier@uni-ulm.de

Torben Widmann
University of Ulm, Germany, torben.widmann@uni-ulm.de

Age Ain't Just a Number: Exploring the Volume vs. Age Dilemma for Textual Data to Enhance Decision Making

Research Paper

Lukas Hägele[1], Mathias Klier[1], Andreas Obermeier[1], and Torben Widmann[1]

[1] University of Ulm, Institute of Business Analytics, Ulm, Germany
{lukas.haegele,mathias.klier,andreas.obermeier,torben.widmann}@uni-ulm.de

Abstract. The common belief that more data leads to better results often leads to all available data being used to derive the best possible decision. However, the age of data can strongly affect data-driven decision making. Consequently, the desire for larger data volume and at the same time contemporary data leads to the "volume vs. age" dilemma, which has not yet been sufficiently researched. In this work, we rigorously investigate the "volume vs. age" dilemma for textual data using four experiments with real-world data containing customer reviews from the Yelp platform. Contributing to theory and practice, we show that more data is not always better, as the effect of data age can outweigh the effect of data volume, resulting in overall poorer performance. Moreover, we demonstrate that different aspects within textual data can exhibit different temporal effects and that considering these effects when selecting training data can clearly outperform existing practices.

Keywords: Data Volume, Data Age, Aspect-based Decision Making, Text Data.

1 Introduction

The rapid proliferation of data and the expanding realm of data analytics have reshaped the landscape of decision making across various domains (Loebbecke and Picot, 2015; Awan et al., 2021; Dubey et al., 2019). In today's data-driven world, organizations recognize the pivotal role of data as a cornerstone for achieving competitive advantage (Hagiu and Wright, 2020; Sandeep et al., 2022; Vassakis et al., 2018). Data fuels the generation of insights along with better products or services, as in the case of the movie streaming platform Netflix, which has achieved tremendous success due to its advanced analytics capabilities and recommender system (Davenport and Harris, 2017; Rataul et al., 2018). Among the different types of data, unstructured textual data play a particularly prominent role, as vast volumes of textual data are generated every day, encompassing sources as diverse as customer reviews, social media, news articles, and more (Egger and Gokce, 2022; Gandomi and Haider, 2015; Weitzenboeck et al., 2022). This influx of unstructured textual data has become a goldmine for data analytics, fostering the development of more sophisticated models and leading to success stories such as

19th International Conference on Wirtschaftsinformatik,
September 2024, Würzburg, Germany

ChatGPT (Kalla and Smith, 2023; Lund and Wang, 2023; Deng and Lin, 2022). A critical success factor for models such as GPT is the vast volume of training data that the algorithms use (Kalla and Smith, 2023; Lund and Wang, 2023), strengthening the common belief in the data-driven paradigm that *more data leads to better results*.

But is this common belief generally valid? Often, all available data is used to derive decisions, without considering the data age, i.e., the time since the data was recorded. However, the quality and validity of data can change over time (Bennin et al., 2020; Spruit and van der Linden, 2019). Such temporal effects can include, for instance, data becoming outdated or changes in the distribution of the input data and their relation to the output data (concept drift) (Agarwal and Nenkova, 2022; Leysen, 2023; Spruit and van der Linden, 2019). An example is the age of customer reviews, which relates to their relevance to the current state of products/services and users and is highly correlated with the helpfulness to users (Luo et al., 2021; Meng et al., 2021). Indeed, data age is a linchpin in the efficacy of data-driven models and can strongly affect derived decisions (Raza and Ding, 2022; Lazaridou et al., 2021; Röttger and Pierrehumbert, 2021). Overall, there exists an inherent contradiction between the desire for larger data volume and the imperative of using contemporary data. We refer to this interplay of the opposing facets, data volume and data age, as the "volume vs. age" dilemma.

Both the influence of data volume and data age have been studied individually in the literature, and insightful results have been obtained for both effects. However, the complex interplay between the need to leverage large data volumes and the imperative of using contemporary data has not been adequately explored – indeed, for the context of textual data, it has not been researched at all. To fill this research gap, we thus investigate the dilemma explicitly for unstructured textual data, using the case of a recommender system based on textual customer reviews. Moreover, we are the first to study the interplay of data volume and data age conducting analyses at the aspect-level – a perspective especially important for textual data. The results reveal that more data is not always better and that handling the "volume vs. age" dilemma for textual data requires a nuanced, aspect-level view, paving the way for more sophisticated training data selection strategies that can significantly outperform existing practices.

2 Related Work

According to the common belief in the data-driven paradigm that *more data leads to better results*, as much (training) data as possible should be used for decision making (Prusa et al., 2015; Lei et al., 2019; Barbedo, 2018). To better understand the impact of data volume on the performance of data-driven methods, many studies have investigated this relationship experimentally. While in some cases, even relatively small data volumes can yield satisfactory results (Fang et al., 2021), the literature generally indicates that increasing data volume further improves performance, while the marginal benefit of using more data decreases (Chen et al., 2017; Sun et al., 2017; Durden et al., 2021). Further literature (Althnian et al., 2021; Langenkämper et al., 2020; Simmonds et al., 2020; Zhu et al., 2016) also emphasizes that more data can lead to better perfor-

mance, while identifying different factors that influence this effect, such as the representation of original distributions, diverse data sources, model complexity, and proper data cleaning. Taken together, these studies provide further insights into the common belief that *more data leads to better results*. They argue that the marginal benefit of more data decreases as data volume increases and that the effect of data volume should be evaluated in conjunction with other critical factors.

Indeed, various data characteristics can have diverse (negative) effects on data-driven decision making (Janssen et al., 2020; Helfert, 2018; Heinrich et al., 2021). One of the characteristics to be considered is data age, which refers to the time since the data was recorded. Due to the corresponding temporal effects such as obsolescence or concept drift, several studies have examined the impact of data age on the performance of data-driven methods. As data can evolve over time, predictions of models using data that is not contemporary anymore experience fluctuations (Kabir et al., 2019; Bennin et al., 2020). Resulting negative effects of data age are particularly evident in unstructured textual data. Several studies concluded from experiments that models perform worse the older the used training data was for cases like recommender systems (Zheng and Horace, 2013), sentiment classification (Lukes and Søgaard, 2018), large language models (Röttger and Pierrehumbert, 2021; Lazaridou et al., 2021), and text classification (Alkhalifa et al., 2023). The studies highlight the critical impact data age can have on the performance of data-driven methods due to changes in the data over time.

Both data volume and data age have emerged as critical drivers of effective decision making. However, the complex interaction of data volume and data characteristics like age, has rarely been addressed in the literature. Only a few studies (e.g., Luca et al., 2022; Roccetti et al., 2019) consider the quality of data in conjunction with data volume, but do not consider data age. Thus, there is paucity of research focusing on the intertwined nature with data age. The study by De Pessemier et al. (2010) stands out as an exception. They particularly investigate recommender system performance when gradually increasing data volume by successively adding data of higher age. They demonstrate that using older data may initially increase performance but can actually decrease recommendation accuracy as the data used is drawn from further back in time.

The “volume vs. age” dilemma has received very little attention in the literature. De Pessemier et al. (2010) examined this dilemma but consider only structured data in terms of user ratings, leaving the impact for unstructured text data unexplored. Furthermore, existing studies that focus on the impact of data age consider temporal effects only in the context of an entire instance, e.g., an entire customer review. As a result, they lack the granularity to differentiate between temporal effects of different features within an instance. For the example of a customer review, the information about the location of a restaurant in a customer review may still be valid while the information about the quality of the food may not be, if these aspects age at different rates (cf. Heinrich and Klier, 2015; Klier et al., 2021). Thus, there is a compelling need to explore the “volume vs. age” dilemma in more detail, focusing explicitly on textual data and taking into account the nuanced effects of different aspects within textual data – both uncharted territory in the current body of literature.

3 Design and Realization of the Experiments

In this section, we present our methodological approach to study the "volume vs. age" dilemma. First, we outline the design of the experiments. Then, we describe the realization of the experiments by elucidating the used textual data and describing the model for data-driven decision making as well as the employed evaluation metrics.

3.1 Experimental Design

In the following, we outline the design of the experiments conducted to investigate the effects of volume and age of textual data. To rigorously study these effects, a data-driven model that constitutes our experimental basis is used to leverage available training data to provide predictions for decision making. In our experiments, we then systematically vary the training data for this model regarding data volume and data age and monitor the resulting performance of the model's predictions on recent and static test data. Our study unfolds across four distinct experiments, each designed to investigate specific facets of the "volume vs. age" dilemma. We commence with two experiments on the effects of data volume and data age, respectively. Then, we elucidate our experimental design for both data volume and data age to capture the "volume vs. age" dilemma. Finally, we conclude with the design of an experiment allowing a more nuanced investigation of the dilemma by differentiating specific aspects within the textual data (e.g., comments regarding *Food* or *Location* in customer reviews).

Experiment (A) – Studying the Data Volume Effect. Our first experiment focuses on the isolated effect of data volume. To study this effect, we iteratively increase the amount of training data (data volume), while the training data is randomly selected from the entire training data corpus to abstract from temporal (data age) effects (Figure 1 (A)). More precisely, we vary the proportion of training data used from the training data corpus in 10% increments from 10% to 100% and evaluate the models' corresponding performance on the test dataset. By selecting the training data randomly, we are able to isolate the volume effect and exclude any potential temporal effects stemming from the age of the training data in a first step.

Experiment (B) – Studying the Data Age Effect. Our second experiment aims to isolate the effect of data age. To systematically analyze this effect, we adopt a controlled approach that has already been used in previous studies (Agarwal and Nenkova, 2022; Röttger and Pierrehumbert, 2021). As illustrated in Figure 1 (B), we construct a series of equally sized training datasets (hereafter referred to as data chunks) from the training data corpus, each containing the same quantity of data instances N_{Chunk} to abstract from potential effects of varying data volume. Moreover, the series of data chunks is constructed to contain training data of increasing age, such that the first data chunk contains the N_{Chunk} most recent and the last chunk the least recent N_{Chunk} data instances of the training data corpus. By selecting data chunks of the same size for training, we are able to isolate the temporal effect and exclude any potential effects due to data volume.

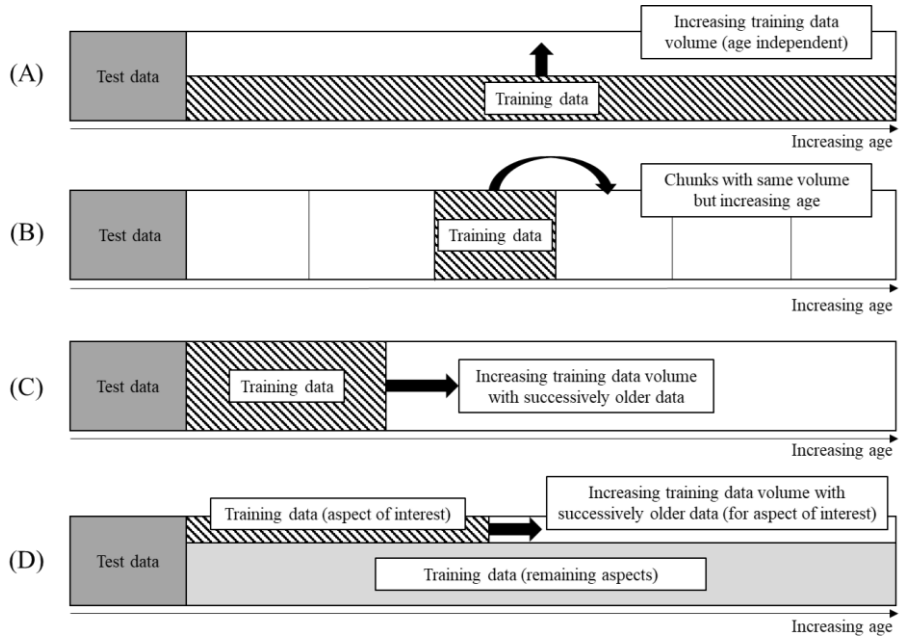

Figure 1. Schematic illustration of the experiments. (A): Increasing data volume by random and age independent sampling. (B): Increasing age of data with same volume. (C): Increasing data by successively adding older data. (D): Increasing data for the aspect of interest.

Experiment (C) – Studying the "Volume vs. Age" Dilemma. With this pivotal experiment, we aim to get an integrated view considering both data volume and data age to be able to rigorously study the "volume vs. age" dilemma (Figure 1 (C)). Thereby, the volume of the training data used is gradually increased by successively adding older data. More precisely, the first training dataset contains the 10% most recent data instances of the training data corpus. Then, the volume of the training data is successively increased by the next 10% most recent remaining data instances. Evaluating the performance of the models trained based on these training datasets on the test dataset, we aim to gain deeper insights into the interplay of data volume and data age, thereby contributing new knowledge regarding the "volume vs. age" dilemma.

Experiment (D) – Studying the "Volume vs. Age" Dilemma at Aspect-Level. According to literature (cf. Heinrich and Klier 2015; Klier et al., 2021), different features of data can age at different rates and their durability is not uniform (cf. Section 2). In the case of customer reviews, for instance, information regarding the aspect *Food* may be subject to different temporal effects than information regarding the aspect *Location*. In our fourth experiment, we aim at more fine-grained insights regarding the interplay of data volume and data age conducting analyses at the aspect-level. More precisely, we conduct Experiment (C) by differentiating specific aspects within the data (e.g., comments regarding *Food* or *Location* in a customer review), as illustrated in Figure 1 (D). Hence, for the aspect under observation, the volume of the training data used is gradually increased by successively adding older data, while the data regarding all

other aspects remains constant (using the whole training data corpus). Thus, we are able to investigate the "volume vs. age" dilemma for each single aspect.

3.2 Realization of the Experiments

Customer reviews serve as paramount example of impactful textual data. By providing valuable, detailed insights into products and services (Mudambi and Schuff, 2010), customer reviews have been shown to support consumers in making purchasing decisions (Dellarocas, 2003; Chevalier and Mayzlin, 2006; Zhu and Zhang, 2010). The contained wealth of information makes them a prime asset for data-driven decision making. Recommender systems are a common and highly relevant field of application leveraging textual data (Kanwal et al., 2021) to derive personalized recommendations (Chen et al., 2015; McAuley and Leskovec, 2013). For the realization of our experiments, we thus use the case of a recommender system for restaurants based on customer reviews.

Dataset. We select a publicly available real-world dataset from Yelp, a leading customer review platform (Yelp Inc., 2023). It contains customer reviews of users sharing their insights about various aspects of restaurants. From this massive dataset, we use customer reviews regarding restaurants in the state of Florida, USA, from the beginning of 2009 to the end of 2019. We choose our test data to contain all customer reviews from the most recent six months, i.e., from July 1st to December 31st, 2019. This test data is static in all our experiments to rigorously compare the performance of our model when the training data is varied. Customer reviews that are not contained in the test data (i.e., those created before July 2019) constitute the whole training data corpus. To enhance the reliability of our analyses, we employ a 10-core test dataset, a well-known preprocessing technique (Cheng et al., 2018; He et al., 2017) so that the dataset exclusively contains customer reviews of users that generated at least ten customer reviews in the test dataset. The resulting dataset comprises 58,128 customer reviews, with a test set containing 9,220 and a training data corpus encompassing 48,908 customer reviews.

Model for Data-Driven Decision Making and Evaluation Metrics. To make recommendations, we use the well-known explicit factor model of Zhang et al. (2014), a common and decisive model with far-reaching impact. To avoid bias from the structured ratings that can influence the effect of the dilemma for textual data, we adapt the model of Zhang et al. (2014) so that the recommendations only depend on the information contained in the customer reviews. The adapted model keeps using the textual customer reviews in the form of aspect-sentiment tuples, e.g., for the exemplary review sentence, *I liked the food, however, the staff was rude*, the resulting aspect-sentiment tuples correspond to (*Food*, positive) and (*Staff*, negative). Hereby, each aspect pertains to a particular feature of the reviewed restaurant and each sentiment reflects the sentiment polarity expressed in the associated text. The aspect-sentiment tuples serve as the foundation for the creation of characteristic profiles for both customers and restaurants. Finally, to yield a recommendation, the model bases its decision on the match between the customer and restaurant profiles. In this line, the model calculates a ranking score, quantifying the match in terms of the similarity between customer and restaurant profiles. To extract the required aspect-sentiment tuples from the customer reviews, we use a state-of-the-art language model for aspect-based sentiment analysis (Yang et al.,

2021) based on transformer models (Devlin et al., 2019). We further employ a standard clustering approach (Mohammed et al., 2020; Reimers and Gurevych, 2019) to group the extracted aspects based on GloVe embedding vectors (Pennington et al., 2014), resulting in a total of 24 distinct aspect classes such as *Food*, *Service*, and *Location* (hereafter just referred to as aspects).

The review-based recommender system generates a ranking score for the relevance of each recommendation (Lü et al., 2012). In line with a widely adopted approach to differentiate between relevant and non-relevant recommendations (Karatzoglou et al., 2013), a restaurant is considered relevant for a specific customer if the associated rating is 3 or higher, while otherwise deemed as non-relevant. To differentiate between both classes, the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve and Mean Average Precision (MAP) have proven to be robust and comprehensive evaluation metrics (cf. Bellogin et al., 2011; Karatzoglou et al., 2013; Ying et al., 2018). The AUC is used to evaluate a recommender system's overall separation performance without directly considering the ranking order of recommendations (Lü et al., 2012). The MAP assesses the precision of a recommender system at various recall levels by considering the presence of relevant restaurants and their positions in the ranking (Yue et al., 2007), thus reasonably complementing the AUC.

4 Results

In this section, we first present the results of the isolated effects of data volume and data age, respectively. Then, we take an integrated view considering both data volume and data age and conclude with our fine-grained analysis differentiating specific aspects within the data.

4.1 The Data Volume Effect

In Experiment (A), we iteratively increase the volume of training data from 10% to 100% of the entire training data corpus in 10% increments, while the respective training data is randomly selected to abstract from temporal effects. The left side of Figure 2 depicts the performance of the trained models on the test datasets with respect to both evaluation metrics employed (AUC and MAP). The results show that increasing data volume goes along with an increase in model performance. Indeed, the performance curves for AUC and MAP show a similar progression and are monotonically increasing. Especially for smaller data volumes, using more data considerably improves model performance. As the data volume continues to increase, the marginal benefit of using more data slightly decreases. Thus, the observed effect that more data increases performance is less evident as more data is available.

4.2 The Data Age Effect

In Experiment (B), we analyze model performance for a series of data chunks containing training data of increasing age but same size to abstract from volume effects. We

chose each data chunk to contain $N_{Chunk} = 6{,}000$ customer reviews. This choice leads to a sufficient number of data chunks to follow the course of the resulting performance curves as data age increases, while ensuring that the number of customer reviews in each chunk is large enough to yield meaningful model outputs. Indeed, model performance shows a negative trend as the age of the training data increases (Figure 2). Thereby, in line with Experiment (A), the performance curves for AUC and MAP exhibit a very similar shape. Despite some fluctuations, the performance curves for both evaluation metrics drop sharply when older data chunks are used.

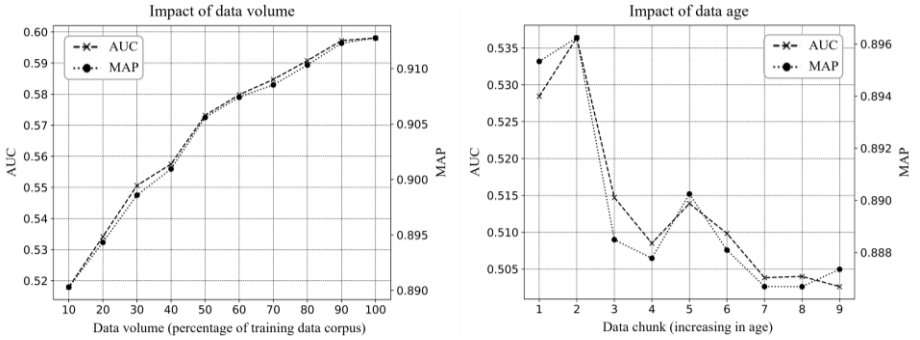

Figure 2. Results of Experiment (A) (left) and Experiment (B) (right) for AUC and MAP

4.3 The "Volume vs. Age" Dilemma

Integrating the effects of data volume and data age, Experiment (C) focuses on the "volume vs. age" dilemma. The results of the experiment are shown in Figure 3 in separate plots for both evaluation measures. By including the curves of Experiment (A), we highlight the difference between increasing the data volume by successively adding older data (Experiment (C)) in contrast to adding randomly selected data (Experiment (A)). The results of Experiment (C) reveal that, for both evaluation metrics, performance initially increases rapidly when increasing data volume by successively adding older data (dashed lines in Figure 3). This trend, however, diminishes quickly as the data volume is further increased with continuously less recent data. In fact, there exists a tipping point when the performance is even decreasing. Thus, when using more and more but less recent data, we do not only observe a reduction in the marginal benefit of increased data volume, but actually a decline in performance. Comparing the results of Experiment (C) with those of Experiment (A) reveals, that except for the last point in each curve (representing the use of all training data and thus coincides), selecting data by age clearly outperforms selecting data randomly. It is noticeable that the performance curve for selecting the data by age starts higher, rises more steeply at the beginning and shows a tipping point that is not observed when selecting data randomly. Based on 20 experimental runs, the Wilcoxon signed-rank test (Wilcoxon, 1945) shows that the best possible performance is significantly higher when selecting training data based on age instead of random ($p < 0.05$ for AUC, $p < 0.01$ for MAP).

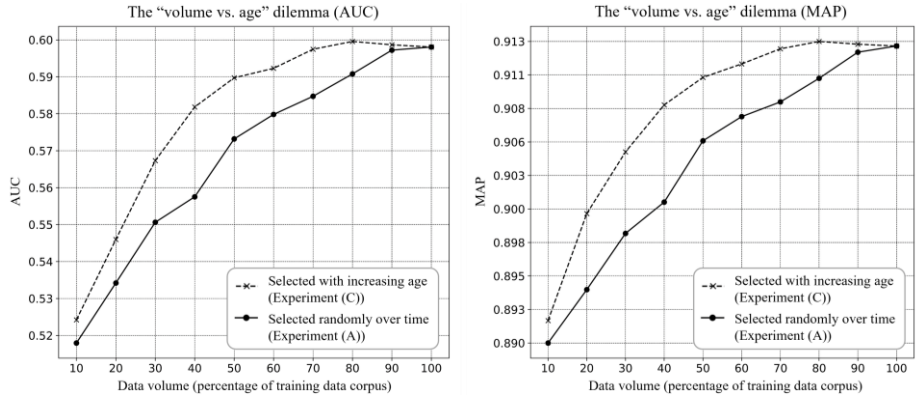

Figure 3. Results of Experiment (C) for AUC (left) and MAP (right)

4.4 The "Volume vs. Age" Dilemma at Aspect Level

To gain more fine-grained insights into the "volume vs. age" dilemma, Experiment (D) differentiates specific aspects within the textual data. Thereby, we consider five (clusters of similar) aspects – *Food*, *Drinks*, *Service*, *Location and Ambience*, and *Price* – and analyze model performance when gradually increasing the volume of the training data used by successively adding older data for the aspect under observation (data for all other aspects is kept constant). As remarkably seen in Figure 4, the respective curves for each aspect differ strongly in their progression and show different trajectories when the associated data volume is increased gradually. For instance, adding more older data for the aspect *Food* decreases the performance rather monotonically. The same holds for the aspect *Price*. On the other hand, for the aspect *Drinks*, the model performance increases as the data volume is increased up to a certain point. The curves for the aspects *Service* and *Location and Ambience* show no clear trends.

Consequently, these in-depth insights regarding the interplay of data volume and data age on the aspect-level give rise to leverage the training data sampling on the aspect-level. In this context, we use the observed results from Experiment (D) and use only the most recent 20 percent of data for *Food* and *Price*, as a strong decrease in model performance is observed when more older data is added for these aspects. For the remaining aspects, we use all data, since for these aspects the addition of more old data does not lead to a clear decrease in model performance. The results using this filtering strategy are also presented in the lower part of Figure 4. The achieved performance when using the aspect-based training data filter noticeably exceeds the performance when using all training data and also exceeds the best performance achieved in Experiment (C), i.e., the optimal trade-off of the "volume vs. age" dilemma for our experiment without aspect-based consideration. To show that the aspect-based selection of data was not a chance hit, we randomly varied the filters for each aspect by ±10% for 20 different experimental runs. Here, the Wilcoxon signed-rank test shows significant outperformance for both evaluation metrics ($p < 0.01$ for AUC and MAP), underscoring the robustness of this aspect-based training data selection.

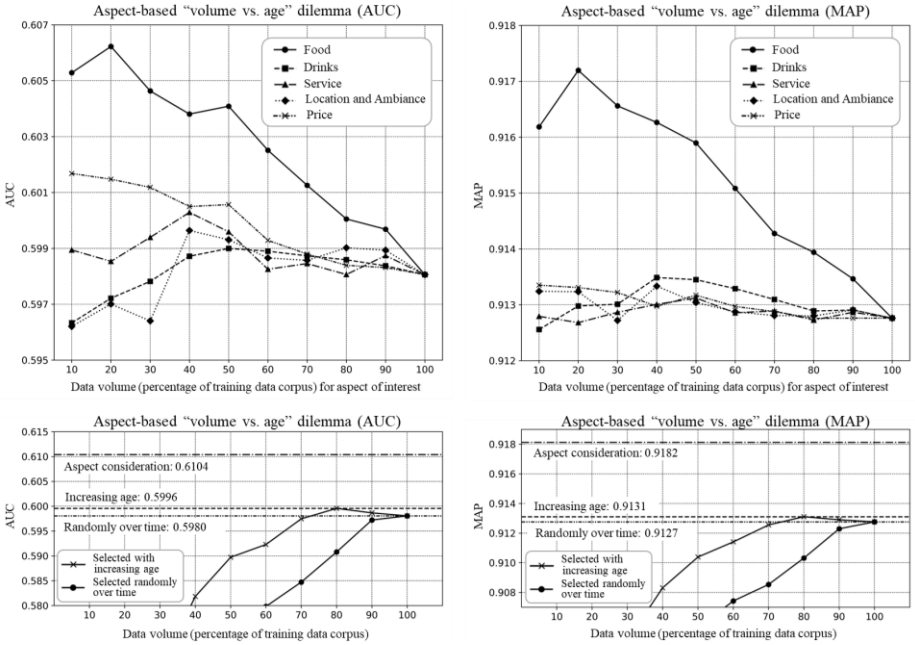

Figure 4. Results of Experiment (D) (upper Figures) and a resulting aspect-based filtering of data (lower Figures) for AUC (left) and MAP (right)

5 Discussion and Conclusion

5.1 Implications for Theory and Practice

Our experiments yield novel insights that challenge conventional assumptions about the advantages of more extensive data volumes in data-driven decision making. Contrary to the common belief that *more data leads to better results*, our findings highlight the interplay between data volume and data age and their impact on the performance of data-driven decision making. Our major contribution is twofold. First, we contribute to theory and practice by rigorously examining the "volume vs. age" dilemma for unstructured data, highlighting that more data is not always better. Second, we are the first to provide in-depth insights regarding the "volume vs. age" dilemma conducting analyses at the aspect-level. The results reveal that handling this dilemma for textual data requires a more differentiated, aspect-level view, paving the way for more sophisticated selection strategies for training data that can significantly outperform existing practices.

Before delving into the "volume vs. age" dilemma, we examine the isolated effects of increasing data volume and increasing data age, respectively. In line with existing literature, we find that excluding any potential temporal effects stemming from data age, more data indeed leads to better results (cf. Barbedo, 2018; Lei et al., 2019) while the marginal benefit of using more data decreases with increasing data volume (cf. Chen et al., 2017; Sun et al., 2017). When isolating the effect of data age, our results reveal

that using older data leads to worse performance (cf. Alkhalifa et al., 2023; Röttger and Pierrehumbert, 2021). Taken together, the Experiments (A) and (B) demonstrate the dichotomy of data volume and data age, resulting in the "volume vs. age" dilemma.

Integrating the effects of data volume and data age, Experiment (C) focuses on the "volume vs. age" dilemma. The results show that more data is not always better, as the age of the data is a very critical factor. In fact, there can exist a tipping point where increasing data volume leads to poorer performance. While performance initially increases as data volume increases by adding less recent data, the marginal benefit decreases. Performance can even decline despite an increase in data volume. Thus, the "volume vs. age" dilemma can not only lead to a reduction of the marginal benefit of increasing data volume but can even cancel it out. Thereby, our experiments show that the diminishing marginal benefit may not be solely due to a saturation regarding data volume – as indicated by Chen et al. (2017) – but also due to the age of the data. Consequently, in some cases, less data can yield better results. For this reason, temporal effects, such as outdated data or concept drift, emerge as critical factors. Thus, practitioners should carefully consider whether less recent data is actually helpful and contributes to performance. Summing up, both facets of the dilemma – data volume and data age – need to be weighted in terms of their potential to improve decision making.

Differentiating specific aspects within the textual data, Experiment (D) allows a more nuanced analysis of the "volume vs. age" dilemma. The results show that aspects within the textual customer reviews can exhibit different effects and influence performance differently. For some aspects, additional data positively contributes to performance while hindering the performance for others. For instance, with respect to the aspect *Location*, performance improves when increasing data volume regarding this aspect. Regarding the aspect *Food*, however, performance decreases when adding less recent data. Thus, our findings advocate for a nuanced consideration of data on individual aspects within textual data. As our study shows, differentiating aspects and incorporating the respective insights into the selection of training data can leverage performance and outperform any data selection strategy not considering the aspect-level.

Different features of data can age at different rates (cf. Heinrich and Klier, 2015; Klier et al., 2021), which is also true for textual data and the underlying aspects. When using textual data, temporal effects should be considered at the aspect-level. A systematic evaluation of aspects – and in particular the effect of their age – can guide decision makers as to whether the performance of data-driven decision making benefits or suffers from the inclusion of more but less recent data. Data selection should therefore be tailored to the specific effects at the aspect-level. Ultimately, our results argue in favor of moving away from undifferentiated data selection strategies and taking a closer look at the underlying data. Indeed, a nuanced, aspect-driven analysis and selection of input data paves the way for improving the performance of data-driven decision making.

5.2 Limitations and Future Work

While our study provides first valuable insights into the "volume vs. age" dilemma, there are limitations that warrant acknowledgement and open avenues for future research. First, this primary exploratory study is dedicated to one application area of data-

driven decision making and is based on a single dataset and a single model. Thus, future research should expand this scope and investigate the "volume vs. age" dilemma for other datasets, data-driven tasks, and models. Moreover, as the observed effects may differ between different application areas, it is crucial to expand the significance and generalizability of our findings by ruling out possible confounding variables and working in a hypothesis-driven manner with different datasets. Second, this study was the first to examine the "volume vs. age" dilemma at a more granular level and found that the five clusters of similar aspects within textual customer reviews have different effects. However, we do not yet differentiate between the individual aspects but considered five clusters. Thus, future research should explore the intricate landscape of textual data by meticulously investigating the "volume vs. age" dilemma at the level of individual aspects to gain insight into even more nuanced effects and ultimately understand how different aspects contribute to overall performance in different contexts.

Other promising directions for future research involve the design of data quality metrics to be able to assess age-related effects of data, such as currency (Klier et al., 2021) and concept drift (Agrahari and Singh, 2022), particularly at the aspect-level of textual data. Integrating such metrics into the decision making calculus of machine learning methods (cf. Hristova, 2014; Firouzian et al., 2019) seems to be a promising way to tackle the "volume vs. age" dilemma. Indeed, by incorporating fine-grained age-related effects, future research has the potential to support data-driven decision making with more robust tools for navigating the dynamics in an ever-changing environment.

5.3 Conclusion

In the field of data-driven decision making, the common belief that *more data leads to better results* usually leads to all available data being used to derive the best possible decisions. However, the age of data can strongly affect the decisions derived from it. Consequently, the desire for larger data volume and at the same time contemporary data leads to the "volume vs. age" dilemma, which has not yet been sufficiently researched. In this paper, we rigorously investigate the "volume vs. age" dilemma for textual data using four experiments with real-world data containing customer reviews from the Yelp platform. We contribute to theory and practice in two ways. First, our results show that more data is not always better. In fact, there can be a tipping point where increasing data volume by adding less recent data worsen the results. Second, we are the first to delve deeper into the "volume vs. age" dilemma not only for textual data but also conducting analyses at the aspect-level. The results show that effectively dealing with the "volume vs. age" dilemma for textual data requires a differentiated view at the aspect-level, which paves the way for more sophisticated training data selection strategies that can significantly outperform existing practices. Overall, the age of data is not just a number, but a critical factor for data-driven decision making.

Acknowledgement: Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 494840328.

References

Agarwal, O. & Nenkova, A. (2022), 'Temporal Effects on Pre-trained Models for Language Processing Tasks', *Transactions of the Association for Computational Linguistics* **10**, 904–921.

Agrahari, S. & Singh, A. K. (2022), 'Concept Drift Detection in Data Stream Mining: A literature review', *Journal of King Saud University - Computer and Information Sciences* **34** (10), 9523–9540.

Alkhalifa, R., Kochkina, E. & Zubiaga, A. (2023), 'Building for tomorrow: Assessing the temporal persistence of text classifiers', *Information Processing & Management* **60** (2), 103200.

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A. B., Alzakari, N., Abou Elwafa, A. & Kurdi, H. (2021), 'Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain', *Applied Sciences* **11** (2), 796.

Awan, U., Shamim, S., Khan, Z., Zia, N. U., Shariq, S. M. & Khan, M. N. (2021), 'Big data analytics capability and decision-making: The role of data-driven insight on circular economy performance', *Technological Forecasting and Social Change* **168**, 120766.

Barbedo, J. G. A. (2018), 'Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification', *Computers and Electronics in Agriculture* **153**, 46–53.

Bellogin, A., Castells, P. & Cantador, I. (2011), Precision-oriented evaluation of recommender systems, *in* 'Proceedings of the fifth ACM conference on Recommender systems', New York, NY, USA: ACM, 333–336.

Bennin, K. E., Ali, N. bin, Borstler, J. & Yu, X. (2020). Revisiting the Impact of Concept Drift on Just-in-Time Quality Assurance, *in* '2020 IEEE 20th International Conference on Software Quality, Reliability and Security (QRS)'. IEEE.

Chen, H., Xiong, F., Wu, D., Zheng, L., Peng, A., Hong, X., Tang, B., Lu, H., Shi, H. & Zheng, H. (2017), Assessing impacts of data volume and data set balance in using deep learning approach to human activity recognition, *in* '2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)', IEEE, 1160–1165.

Chen, L., Chen, G. & Wang, F. (2015), 'Recommender systems based on user reviews: the state of the art', *User Modeling and User-Adapted Interaction* **25** (2), 99–154.

Cheng, Z., Ding, Y., Zhu, L. & Kankanhalli, M. (2018), Aspect-Aware Latent Factor Model, *in* 'Proceedings of the 2018 World Wide Web Conference', 639–648.

Chevalier, J. A. & Mayzlin, D. (2006), 'The Effect of Word of Mouth on Sales: Online Book Reviews', *Journal of Marketing Research* **43** (3), 345–354.

Davenport, T. & Harris, J. (2017), *Competing on Analytics: Updated, with a New Introduction. The New Science of Winning*. Harvard Business Press.

De Pessemier, T., Dooms, S., Deryckere, T. & Martens, L. (2010), Time dependency of data quality for collaborative filtering algorithms, *in* 'Proceedings of the fourth ACM conference on Recommender systems', New York, NY, USA: ACM, 281–284.

Dellarocas, C. (2003), 'The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms', *Management Science* **49** (10), 1407–1424.

Deng, J. & Lin, Y. (2022), 'The Benefits and Challenges of ChatGPT: An Overview', *Frontiers in Computing and Intelligent Systems* **2** (2), 81–83.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *in* 'Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) 2019', 4171–4186.

Dubey, R., Gunasekaran, A., Childe, S. J., Papadopoulos, T., Luo, Z., Wamba, S. F. & Roubaud, D. (2019), 'Can big data and predictive analytics improve social and environmental sustainability?', *Technological Forecasting and Social Change* **144**, 534–545.

Durden, J. M., Hosking, B., Bett, B. J., Cline, D. & Ruhl, H. A. (2021), 'Automated classification of fauna in seabed photographs: The impact of training and validation dataset size, with considerations for the class imbalance', *Progress in Oceanography* **196**, 102612.

Egger, R. & Gokce, E. (2022), Natural Language Processing (NLP): An Introduction, *in* 'Applied Data Science in Tourism', 307–334, Springer International Publishing.

Fang, Y., Wang, J., Ou, X., Ying, H., Hu, C., Zhang, Z. & Hu, W. (2021), 'The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients', *Physics in medicine and biology* **66** (18), 185012.

Firouzian, I., Zahedi, M. & Hassanpour, H. (2019), 'Investigation of the Effect of Concept Drift on Data-Aware Remaining Time Prediction of Business Processes', *International Journal of Nonlinear Analysis and Applications* **10** (2), 153–166.

Gandomi, A. & Haider, M. (2015), 'Beyond the hype: Big data concepts, methods, and analytics', *International Journal of Information Management* **35** (2), 137–144.

Gupta, M., Akiri, C., Aryal, K., Parker, E. & Praharaj, L. (2023), 'From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy', *IEEE Access* **11**, 80218–80245.

Hagiu, A. & Wright, J. (2020), 'When Data Creates Competitive Advantage', *Harvard Business Review* **98**, 94–101.

He, X., Liao, L., Zhang, H., Nie, L., Hu, X. & Chua, T.-S. (2017), Neural Collaborative Filtering, *in* 'Proceedings of the 26th International Conference on World Wide Web', Perth, Australia.

Heinrich, B., Hopf, M., Lohninger, D., Schiller, A. & Szubartowicz, M. (2021), 'Data quality in recommender systems: the impact of completeness of item content data on prediction accuracy of recommender systems', *Electronic Markets* **31** (2), 389–409.

Heinrich, B. & Klier, M. (2015), 'Metric-Based Data Quality Assessment — Developing and Evaluating a Probability-Based Currency Metric', *Decision Support Systems* **72**, 82–96.

Helfert, M. (2018), 'Perspectives of Big Data Quality in Smart Service Ecosystems (Quality of Design and Quality of Conformance)', *Journal of Information Technology Management* **10** (4), 72–83.

Hristova, D. (2014), Considering Currency in Decision Trees in the Context of Big Data, *in* 'International Conference on Interaction Sciences'.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S. & Janowski, T. (2020), 'Data governance: Organizing data for trustworthy Artificial Intelligence', Government Information Quarterly **37** (3).

Kabir, M. A., Keung, J. W., Bennin, K. E. & Zhang, M. (2019), Assessing the Significant Impact of Concept Drift in Software Defect Prediction, *in* '2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)', IEEE, 53–58.

Kalla, D. & Smith, N. (2023), 'Study and Analysis of Chat GPT and its Impact on Different Fields of Study', *International Journal of Innovative Science and Research Technology* **8** (3), 827–833.

Kanwal, S., Nawaz, S., Malik, M. K. & Nawaz, Z. (2021), 'A Review of Text-Based Recommendation Systems', *IEEE Access* **9**, 31638–31661.

Karatzoglou, A., Baltrunas, L. & Shi, Y. (2013), Learning to rank for recommender systems, *in* 'Proceedings of the 7th ACM conference on Recommender systems', New York, NY, USA: ACM, 493–494.

Klier, M., Moestue, L., Obermeier, A. & Widmann, T. (2021), Event-Driven Assessment of Currency of Wiki Articles: A Novel Probability-Based Metric, *in* 'International Conference on Information Systems', Austin, TX.

Langenkämper, D., van Kevelaer, R., Purser, A. & Nattkemper, T. W. (2020), 'Gear-Induced Concept Drift in Marine Images and Its Effect on Deep Learning Classification', *Frontiers in Marine Science* **7**, 538862.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., C. de Masson d'Autume, Kocisky, T., Ruder, S., Yogatama, D., Cao, K., Young, S. & Blunsom, P. (2021), 'Mind the gap: Assessing temporal generalization in neural language models', *in* 'Advances in Neural Information Processing Systems' (NeurIPS), Virtual Event.

Lei, S., Zhang, H. & Su, S. (2019), How training data affect the accuracy and robustness of neural networks for image classification, *in* ICLR, New Orleans, LA, USA.

Leysen, J. (2023). Exploring Unlearning Methods to Ensure the Privacy, Security, and Usability of Recommender Systems, *in* 'Proceedings of the 17th ACM Conference on Recommender Systems', New York, NY, USA: ACM.

Loebbecke, C. & Picot, A. (2015), 'Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda', *The Journal of Strategic Information Systems* **24** (3), 149–157.

Lü, L., Medo, M., Yeung, C. H., Zhang, Y.-C., Zhang, Z.-K. & Zhou, T. (2012), 'Recommender systems', *Physics Reports* **519** (1), 1–49.

Luca, A. R., Ursuleanu, T. F., Gheorghe, L., Grigorovici, R., Iancu, S., Hlusneac, M. & Grigorovici, A.: (2022), 'Impact of quality, type and volume of data used by deep learning models in the analysis of medical images', *Informatics in Medicine Unlocked* **29**, 100911.

Lukes, J. & Søgaard, A. (2018), Sentiment analysis under temporal shift, *in* 'Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis', Stroudsburg, PA, USA: Association for Computational Linguistics, 65–71.

Lund, B. D. & Wang, T. (2023), 'Chatting about ChatGPT: how may AI and GPT impact academia and libraries?', *Library Hi Tech News* **40** (3), 26–29.

Luo, L., Duan, S., Shang, S. & Pan, Y. (2021), 'What makes a helpful online review? Empirical evidence on the effects of review and reviewer characteristics', *Online Information Review* **45** (3), 614–632.

McAuley, J. & Leskovec, J. (2013), Hidden factors and hidden topics, *in* 'Proceedings of the 7th ACM conference on Recommender systems', New York, NY, USA: ACM.

Meng, Y., Yang, N., Qian, Z. & Zhang, G. (2021), 'What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values', *Journal of Theoretical and Applied Electronic Commerce Research* **16** (3), 466–490.

Mohammed, S. M., Jacksi, K. & Zeebaree, S. R. M. (2020), Glove Word Embedding and DBSCAN algorithms for Semantic Document Clustering, *in* '2020 International Conference on Advanced Science and Engineering (ICOASE)', IEEE, 1–6.

Mudambi, S. M. & Schuffm, D. (2010), 'What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com', *MIS Quarterly* **34** (1), 185–200.

Pennington, J., Socher, R. & Manning, C. (2014), Glove: Global Vectors for Word Representation, *in* 'Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP): Association for Computational Linguistics', 1532–1543.

Prusa, J., Khoshgoftaar, T. M. & Seliya, N. (2015), The Effect of Dataset Size on Training Tweet Sentiment Classifiers, *in* '2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)', IEEE, 96–102.

Rataul, P., Tisch, D. G. & Peter, Z. (2018), *Netflix. Dynamic capabilities for global success*. SAGE Publications: SAGE Business Cases Originals.

Ray, P. P. (2023), 'ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope', *Internet of Things and Cyber-Physical Systems* **3**, 121–154.

Raza, S. & Ding, C. (2022), 'News recommender system: a review of recent progress, challenges, and opportunities', *Artificial Intelligence Review* **55** (1), 749–800.

Reimers, N. & Gurevych, I. (2019), Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks, *in* 'Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing', Hong Kong, China, 3982–3992.

Roccetti, M., Delnevo, G., Casini, L. & Cappiello, G. (2019), 'Is bigger always better? A controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures', *Journal of Big Data* **6** (1), 1–23.

Röttger, P. & Pierrehumbert, J. (2021), Temporal Adaptation of BERT and Performance on Downstream Document Classification: Insights from Social Media, *in* 'Findings of the Association for Computational Linguistics: EMNLP 2021', Stroudsburg, PA, USA: Association for Computational Linguistics, 2400–2412.

Sahoo, N., Singh, P. V. & Mukhopadhyay, T. (2012), 'A Hidden Markov Model for Collaborative Filtering', *MIS Quarterly* **36** (4), 1329–1356.

Sandeep, S. R., Ahamad, S., Saxena, D., Srivastava, K., Jaiswal, S. & Bora, A. (2022), 'To understand the relationship between Machine learning and Artificial intelligence in large and diversified business organisations', *Materials Today: Proceedings* **56** (4), 2082–2086.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B. & O'Hara, R. B. (2020), 'Is more data always better? A simulation study of benefits and limitations of integrated distribution models', *Ecography* **43** (10), 1413–1422.

Spruit, M. & van der Linden, V. (2019), 'BIDQI: The Business Impacts of Data Quality Interdependencies Model', *Technical Report Series* UU-CS-2019 (001), 1–25.

Sun, C., Shrivastava, A., Singh, S. & Gupta, A. (2017), Revisiting Unreasonable Effectiveness of Data in Deep Learning Era, *in* '2017 IEEE International Conference on Computer Vision (ICCV)', IEEE, 843–852.

Vassakis, K., Petrakis, E. & Kopanakis, I. (2018), Big Data Analytics: Applications, Prospects and Challenges, *in* 'Mobile Big Data – A Roadmap from Models to Technologies', 3–20. Cham: Springer.

Weitzenboeck, E. M., Lison, P., Cyndecka, M. & Langford, M. (2022), 'The GDPR and unstructured data: is anonymization possible?', *International Data Privacy Law* **12** (3), 184–206.

Wilcoxon, F. (1945), 'Individual Comparisons by Ranking Methods', *Biometrics Bulletin* **1** (6)

Yang, H., Zeng, B., Yang, J., Song, Y. & Xu, R. (2021), 'A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction', *Neurocomputing* **419**, 344–356.

Yelp Inc. (2023), Yelp Open Dataset, URL: https://www.yelp.com/dataset. Accessed: 12.03.2024.

Ying, H., Zhuang, F., Zhang, F., Liu, Y., Xu, G., Xie, X., Xiong, H. & Wu, J. (2018), Sequential recommender system based on hierarchical attention network, *in* 'IJCAI International Joint Conference on Artificial Intelligence', 3926–3932.

Yue, Y., Finley, T., Radlinski, F. & Joachims, T. (2007), A support vector method for optimizing average precision, *in* 'Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval', New York, NY, USA: ACM.

Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y. & Ma, S. (2014), Explicit Factor Models for Explainable Recommendation based on Phrase-level Sentiment Analysis, *in* 'Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval', 83–92.

Zheng, G., Horace, HS. Ip. (2013), Effectiveness of the data generated on different time in latent factor model, *in* 'Proceedings of the 7th ACM conference on Recommender systems', 327-330.

Zhu, F. & Zhang, X. (2010), Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics, *Journal of Marketing* **74** (2), 133–148.
Zhu, X., Vondrick, C., Fowlkes, C. C. & Ramanan, D. (2016), Do We Need More Training Data?, *International Journal of Computer Vision* **119** (1), 76–92.