Model Hubs and Beyond: Analyzing Model Popularity, Performance, and Documentation

Pritam Kadasi[1], Sriman Reddy Kondam[1], Srivathsa Vamsi Chaturvedula[1], Rudranshu Sen[*2], Agnish Saha[*2], Soumavo Sikdar[*2], Sayani Sarkar[*2], Suhani Mittal[4†], Rohit Jindal[3], Mayank Singh[1]

[1]Indian Institute of Technology Gandhinagar, Gujarat, India
[2]Jadavpur University, West Bengal, India,
[3]Vellore Institute of Technology, Tamil Nadu, India
[4]Citicorp Services India Private Limited, Pune, Maharashtra, India
{pritam.k, singh.mayank}@iitgn.ac.in

Abstract

With the massive surge in ML models on platforms like Hugging Face, users often lose track and struggle to choose the best model for their downstream tasks, frequently relying on model popularity indicated by download counts, likes, or recency. We investigate whether this popularity aligns with actual model performance and how the comprehensiveness of model documentation correlates with both popularity and performance. In our study, we evaluated a comprehensive set of 500 Sentiment Analysis models on Hugging Face. This evaluation involved massive annotation efforts, with human annotators completing nearly 80,000 annotations, alongside extensive model training and evaluation. Our findings reveal that model popularity does not necessarily correlate with performance. Additionally, we identify critical inconsistencies in model card reporting: approximately 80% of the models analyzed lack detailed information about the model, training, and evaluation processes. Furthermore, about 88% of model authors overstate their models' performance in the model cards. Based on our findings, we provide a checklist of guidelines for users to choose good models for downstream tasks.

Datasets — https://github.com/pskadasi/model-hubs

1 Introduction

In the aftermath of the groundbreaking Transformers paper (Vaswani et al. 2017), the field of Natural Language Processing (NLP) has undergone a rapid evolution, witnessing an unprecedented surge in research and development. This surge has given rise to a diverse array of models spanning various AI domains, including both vision (Khan et al. 2022) and NLP (Wolf et al. 2020) tasks. To facilitate the dissemination and utilization of these models, researchers have established "Model Hubs" – platforms designed to curate and categorize collections of open models and datasets, fostering sharing and collaboration for a wide range of open models. Among these hubs, the Hugging Face (HF)[1] stands out as the largest model repository on the internet, boasting over 930K models as of September 10th, 2024 and expected to surpass 1M mark[2].

This surge in model availability can be attributed to the influence of social media hype, leading to a significant increase in model downloads/likes. However, this phenomenon resembles the "rich get richer" dynamic (Barabási and Albert 1999), where popular[3] models attract more attention and downloads, perpetuating their dominance. Despite the widespread adoption of these models, users often engage with them in downstream tasks without conducting rigorous verification and validation, raising concerns about their reliability and performance (Jiang et al. 2023). To investigate this, we evaluate models with high download and like counts for a specific task – "Sentiment Analysis," on HF. We expand our evaluation to include analysis on different dataset difficulty categories (Swayamdipta et al. 2020). Our goal is to explore how models obtained from HF perform across various dataset difficulty categories, considering that users may utilize these downloaded models in different contexts.

Furthermore, in response to the growing need for transparency, accountability, and reproducible NLP research, researchers have introduced datasheets (Gebru et al. 2021) and model cards (Mitchell et al. 2019) as mechanisms to provide detailed insights into model characteristics and associated datasets. However, despite the availability of model cards and datasheets as a platform for authors to furnish comprehensive information about their models and datasets, a sub-

[*]These authors contributed equally.

[†]Work done while at IIT Gandhinagar as part of B.Tech project.

[1]https://huggingface.co/

[2]https://x.com/ClementDelangue/status/1674709427557998593?s=20

[3]Throughout the rest of the manuscript, we commonly refer to models as "popular" based on their #downloads and #likes. Essentially, the more #downloads and #likes a model has, the more popular it is perceived to be.

stantial number of authors fail to provide sufficient details, leading to a lack of transparency and trustworthiness (Singh et al. 2023; Liu et al. 2024). The absence of crucial details in model cards not only hampers reproducibility but also hinders the progress of NLP research. To address these concerns, we conduct a thorough manual inspection of model cards for models with high download/like counts on HF, uncovering instances of missing information. We further extend this investigation to understand the impact of missing information in model cards on their download/like count and to determine its correlation with the model's performance.

Given the prevalence of models on platforms like HF, it is pertinent to investigate whether these platforms serve merely as dumping grounds[4][5] for models or if they genuinely contribute to advancing NLP research. Our investigation aims to shed light on these challenges and prompt discussions on the importance of transparent documentation, open models, and reproducible research practices in the NLP.

Overall, our contributions can be summarized as follows:

1. We thoroughly evaluate the 500 most downloaded and liked models, specifically the top-ranked models based on the #downloads, which we refer to as M_{500}, for the Sentiment Analysis (SA) task from HF to determine if high popularity, as represented by #downloads and #likes, actually guarantees better model performance. Our key findings reveal that model performance showed a weak positive correlation and, in some cases, even exhibited a negative correlation with download and like counts.
2. Further, we conduct a large-scale manual study to verify the completeness of information in Model Cards for M_{500} by carefully annotating every element in each section, requiring massive annotation efforts. Our investigation reveals that more than half of the M_{500} lack key details about the training process, and a striking 96% neglect to provide any information about critical aspects such as Bias, Risks, and Limitations. We observe a high popularity of highly documented models and vice versa.
3. To the best of our knowledge, this is the first large-scale exhaustive study conducted on models from the HF platform. Additionally, we create a novel Reddit SA dataset to evaluate the performance of top models on a completely unseen dataset. The results reveal poor generalization performance and a weak positive correlation with popularity.

2 Related Works

Prior studies have examined various aspects of the HF platform, including pre-trained models (PTMs) reuse and popularity's influence on adoption (Jiang et al. 2023; Taraghi et al. 2024). Other works have assessed platform vulnerabilities (Kathikar et al. 2023), environmental impact (Castaño et al. 2023), and ML model maintenance (Castaño et al. 2023). Additionally, efforts have been made to categorize PTMs for software engineering tasks, such as automatic classification using a public HF dump (Sipio et al. 2024). In contrast, our study investigates the relationship between model popularity and performance on HF, questioning whether high popularity guarantees better performance.

Related works (Mitchell et al. 2019; Crisan et al. 2022) have proposed model cards for experts and non-experts, while (Liu et al. 2024) addresses information incompleteness in model cards and proposes an automated generation approach using LLMs. Another study (Pepe et al. 2024) has investigated the transparency of pre-trained transformer models on HF, highlighting limited exposure of training datasets and biases. Our work differs by examining how information incompleteness in model cards influences model popularity and performance. In summary, our research evaluates popular HF models, exploring their relationship with model card completeness and performance.

3 Experiments

3.1 Models Filtering Criteria

We filter the most popular models on HF focusing on text classification models within NLP due to their abundance[6]. Specifically, within the text classification task, we opt for SA as a representative use case because of its extensive exploration in both industry and academia. We use HFApi[7] to extract metadata for all text classification models, searching for the keyword "sentiment" in their model names. During our exploration of HF platform, we identify models like `ProsusAI/finbert`, which, despite lacking the explicit "sentiment" keyword, are SA models. To cover such cases comprehensively, we broaden our search (using HFApi) to models without the sentiment keyword. Within these models, we identify SA models by manually examining their output labels (refer to §A.1). For our experiments, we confine our analysis to three-class SA models to maintain simplicity and avoid complexity introduced by models with more than three class labels. Following this process, we identify and include 1207 models[8]that meet our criteria (refer to §A.1 for more details).

Continuing with our analysis, we employ a ranking system r based on metrics such as #downloads, #likes, and the date (last modified) of the model. We determine the rank r based on these metrics. Specifically, for downloads and likes, we calculate r_d and r_l by arranging the number of downloads and likes in descending order and assigning ranks from 1 to 1207. Similarly, for the date, we calculate r_{da} by arranging the dates in descending order, ensuring that more recent dates receive higher ranks.

Subsequently, we select the top 500 models based on r_d, out of the three possible ranks (r_d, r_l, and r_{da}) due to its wide adoption (Jiang et al. 2023), forming the collection M_{500} (refer to Table 7 in §A). We select the top 500 models for our analysis, as the first 100 alone account for over 99.99% of the total downloads among the 1207 SA models. While these top models dominate usage, extending the scope

[4]https://x.com/sd_marlow/status/1674967110739435525?s=20
[5]https://x.com/KastelBert/status/1674761863798693892?s=20

[6]https://huggingface.co/tasks
[7]https://huggingface.co/docs/huggingface_hub/package_reference/hf_api
[8]These are the models filtered as of August 26th, 2024

to 500 models allowed for broader coverage. Models beyond this threshold have extremely low download counts, following a power-law distribution, and thus contribute minimally to the NLP community. Therefore, we restrict our analysis to the top 500 models.

3.2 Model Card Inspection

We also conduct a manual inspection of M_{500} to identify any discrepancies and incompleteness of information in their model cards (see Figure 1). Model cards (Mitchell et al. 2019) are short documents that accompany ML models, furnishing details about the models, it's usage contexts, performance evaluation procedures, and other pertinent information. To assess the comprehensiveness of information within these cards, four annotators manually review each model's model card on HF (refer to §B for annotator details). They annotate each of the 39 elements in a model card across sections, such as "Training Data", "Training Procedure" and "Training Hyperparameters" in the **Training Details** section, by referring to the HF model card template[9]. Overall, they annotate approximately 78,000 elements (500 models × 39 elements × 4 annotators). Each element is categorized as Full Information Available (FIA), Partial Information Available (PIA), or No Information Available (NIA). For instance, if a model card section contains all the required information according to the template, it is labeled as FIA. If an annotator needs to search for information, such as referring to the research paper URL provided instead of the requested information, or if the information is not provided exactly as per the template or is located in other sections, it is categorized as PIA. If no information is available, it is labeled as NIA. Instances where all annotators disagree result in a "U" label, indicating uncertainty.

To illustrate the labeling process in detail, we use the model card for `distilbert/distilgpt2`[10] as an ideal example for comparison, given that it contains nearly all relevant details. For the **Training Details** section of this model card, we label "Training Data" and "Training Procedure" as FIA, since comprehensive information is explicitly provided. However, for "Training Hyperparameters", we label NIA, as there is no explicit mention of this information in the model card.

In contrast, consider an example from M_{500}: `philschmid/distilbert-base-multilingual-cased-sentiment`[11]. For the **Training Details** section, an annotator labels "Training Data" as PIA, since there is no direct mention of the training data in the "Training Data" element of the model card section. However, relevant information about the dataset (`amazon_multi_reviews`[12]) is provided in the model description. For "Training Procedure", the label is NIA because no related information is stated, while "Training Hyperparameters" is labeled as FIA, as all details regarding hyperparameters are explicitly mentioned in the model card.

Similarly, annotators systematically label all sections of model cards. The final label for each element is determined by majority vote among annotators. These final labels are then used to calculate the percentages as displayed in Figures 3 and 5.

3.3 Datasets

We utilize three distinct datasets ($\mathcal{D}$): TweetEval[13] ($\mathcal{D}_t$) (Barbieri et al. 2020), Amazon Multi Reviews[14] ($\mathcal{D}_a$) (Keung et al. 2020), and Financial Phrasebank[15] ($\mathcal{D}_f$) (Malo et al. 2014). Notably, these datasets hold widespread recognition and are commonly employed as benchmark datasets (Barbieri et al. 2020; Zhang et al. 2023; Shah et al. 2022; Wang, Yang, and Wang 2023; Hussain et al. 2021; Kurihara, Kawahara, and Shibata 2022). Furthermore, approximately 72% of the SA models within M_{500} are originally trained on these datasets, among those in M_{500} that disclosed their training datasets (around 15.4%,) which is around 55 models. While $\mathcal{D}_t$ and $\mathcal{D}_f$ are representative of three-class SA datasets, $\mathcal{D}_a$ initially comprises a five-class (0–4) SA classification schema. To bring uniformity in our study, we transform $\mathcal{D}_a$ into a three-class SA dataset by excluding instances with labels 1 and 3. Further details regarding these datasets are provided in Table 1.

3.4 Training Strategies

We explore two distinct strategies for evaluating M_{500} as outlined below:

Evaluate In this strategy, we assess the performance of M_{500} models without any prior training on the datasets $\mathcal{D}$.

Finetune In this strategy, we finetune M_{500} on the train set and evaluate its performance on the corresponding test set of $\mathcal{D}$.

3.5 Training and Evaluation Methodology

We thoroughly evaluate the M_{500} models to assess their performance across different categories of dataset difficulty (Easy, Ambiguous, and Hard) using dataset cartography (Swayamdipta et al. 2020). This evaluation helps us understand how these models perform across various datasets based on their difficulty levels, given that models can be used in downstream tasks involving datasets of varying difficulty.

Dataset cartography (Swayamdipta et al. 2020; Kadasi and Singh 2023) framework captures training dynamics during training to categorize instances into easy, ambiguous,

[9]https://huggingface.co/docs/hub/model-card-annotated
[10]https://huggingface.co/distilbert/distilgpt2
[11]https://huggingface.co/philschmid/distilbert-base-multilingual-cased-sentiment
[12]https://huggingface.co/datasets/mteb/amazon_reviews_multi
[13]https://huggingface.co/datasets/cardiffnlp/tweet_eval
[14]https://huggingface.co/datasets/defunct-datasets/amazon_reviews_multi
[15]https://huggingface.co/datasets/takala/financial_phrasebank
[16]As of February 6th, 2024, sourced from HF datasets webpage. Additionally, $\mathcal{D}_a$ is sourced from https://huggingface.co/datasets/mteb/amazon_reviews_multi, since the official dataset has been taken down. For $\mathcal{D}_f$, we split the data in the ratio of 70:10:20 for train, val and test sets repectively.

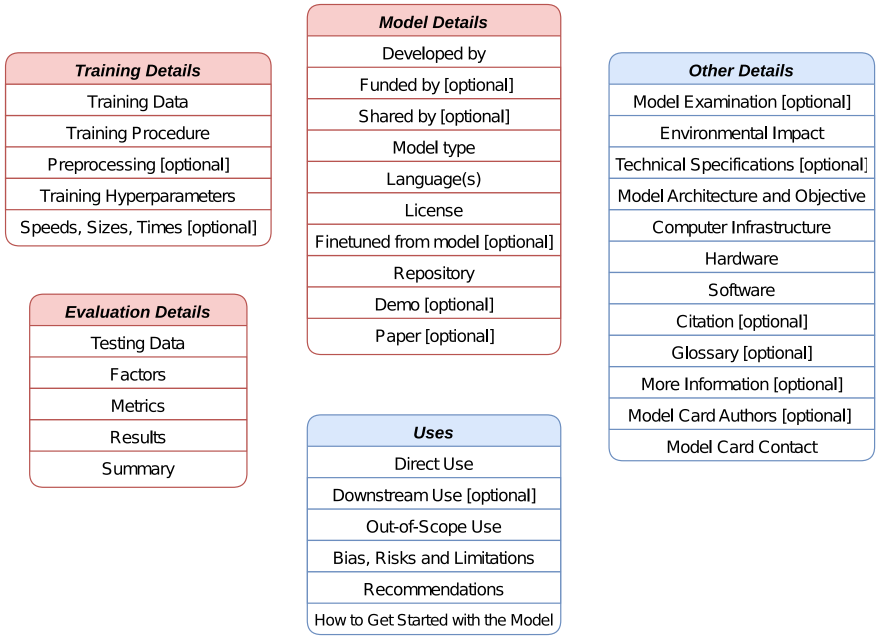

Figure 1: An illustration of a Model Card

Dataset	#Instances (Train/Val/Test)	#Class Labels	#Models Trained	#Downloads All Time	#Instances (Easy/Ambi/Hard)
$\mathcal{D}_t$	45,615/12,284/2,000	3	620	2,057,849	35,793/16,013/8,093
$\mathcal{D}_a$	120,000/3,000/3,000	3	152	219,774	105,052/12,479/8,469
$\mathcal{D}_f$	4,846/-/-	3	130	462,906	4,682/163/1

Table 1: Dataset Statistics[16]

		$\mathcal{D}_t$			$\mathcal{D}_a$			$\mathcal{D}_f$
		Easy	Ambi	Hard	Easy	Ambi	Hard	Easy
Evaluate	r_d	0.270*	0.241*	0.085	0.255*	0.243*	0.241*	0.196*
	r_l	0.164*	0.130*	0.071	0.136*	0.147*	0.156*	0.187*
	r_{da}	-0.136*	-0.126*	-0.018	-0.087	-0.083	-0.125*	-0.098*
Finetune	r_d	-0.089	-0.145*	0.147*	-0.125*	-0.142*	0.206*	-0.069
	r_l	-0.143*	-0.165*	0.161*	-0.177*	-0.181*	0.259*	-0.074
	r_{da}	0.019	0.016	-0.080	0.001	-0.008	-0.054	-0.010

Table 2: Correlation between r and the performance (F1) of models in M_{500} across different dataset categories. Values highlighted with * represent statistical significant values with p-value < 0.05. For $\mathcal{D}_f$, we did not perform any evaluation strategy for ambiguous and hard instances, since the #instances categorized as ambiguous and hard is very low, which is insufficient for training and evaluation, therefore values are not calculated and displayed.

and hard instances. It categorizes the instances based on confidence measured as the mean probability of the correct label across epochs, and variability, represented by the variance of the aforementioned confidence. This analysis creates the datamaps as displayed in Figure 16, where it has three distinct regions: easy-to-learn (**Easy**), ambiguous-to-learn (**Ambiguous**), and hard-to-learn (**Hard**) for the model.

We train the RoBERTa[17] (Liu et al. 2019) model, capturing training dynamics through dataset cartography on $\mathcal{D}$, which comprises train, valid, and test splits (refer to Table 1 for #instances categorized as easy, ambiguous and hard). Next, we apply Evaluate and Finetune strategies (see §3.4) to the test splits of instance difficulty categories.

Before conducting experiments, we perform hyperparameter tuning on the RoBERTa model for $\mathcal{D}$ to identify optimal hyperparameters (refer to §C for more details). These optimal hyperparameters are then consistently used across all experiments.

Evaluation Metrics To present our results effectively, we calculate F1 scores. Additionally, we calculate Spearman rank correlation (Spearman 1904) to examine the relationship between model performance and their corresponding r for $\mathcal{D}$ across different difficulty categories. We use the following terminology in Table 3 throughout the rest of the paper.

Strength	Coefficient Interval
Very Weak	0.00 - 0.199
Weak	0.20 - 0.399
Medium	0.40 - 0.599
Strong	0.60 - 0.799
Very Strong	0.80 - 1.000

Table 3: Summary of Spearman correlation coefficient strengths

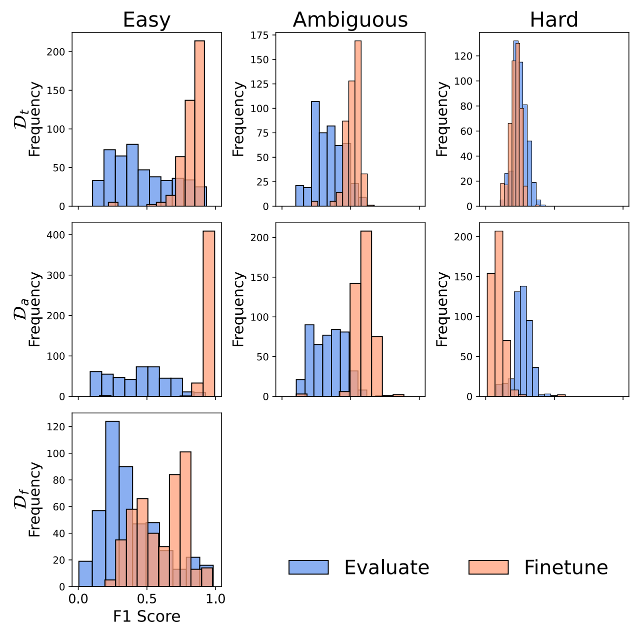

Figure 2: Distribution of F1 scores for models in M_{500} across different difficulty categories in dataset $\mathcal{D}$, under both the Evaluate and Finetune strategies.

4 Results and Discussions

4.1 Does High Popularity Necessarily Correlate With Better Model Performance?

With reference to Table 2, for $\mathcal{D}_t$, we observe weak to very weak correlations between model performance on easy and ambiguous instances and the metrics r_d, r_l, and r_{da}. This indicates that for this dataset, models highly ranked in terms of r_d and r_l and low ranked in terms of r_{da} (older models) are likely to perform better. For hard instances, we do not draw any conclusions since the scores are statistically insignificant.

This weak correlation is due to models that are highly ranked with respect to r_d perform pooorer as compared to some of the models that are low ranked (e.g., `oliverguhr/german-sentiment-bert`[18], r_d: 12, f1-rank[19]: 266 on $\mathcal{D}_t$ (easy)). This trend could be attributed to "popularity bias". Models that are popular (highly ranked) might be favored due to their visibility regardless of their true performance. Community preferences, familiarity with certain model architectures, or advertising on social media by model card authors could play a significant role. Some users may prioritize factors like ease of integration, pretraining details, or compatibility over actual downstream performance.

Model downloads may also reflect the utility of a model for specific tasks rather than its overall performance. For instance, a model might perform exceptionally well for a specialized task but is still broadly downloaded because it fits specific user needs (e.g., `ProsusAI/finbert`[20], r_d: 5). Conversely, a high-performing model on certain benchmarks may not meet the requirements of a large portion of users, leading to fewer downloads (e.g., `cardiffnlp/xlm-v-base-tweet-sentiment-ar`[21], r_d: 395).

After applying the fine-tuning strategy, most models that performed poorly in the Evaluate strategy showed performance gains, resulting in a shift from very weak positive to very weak negative correlation with r_d and r_l across all difficulty categories.

For $\mathcal{D}_a$, similar to $\mathcal{D}_t$, we observe weak to very weak correlations between model performance and the metrics r. In the fine-tuning strategy as depicted in Table 2, we observe a very weak negative correlation. Here, fine-tuning models has degraded the correlation further. For $\mathcal{D}_f$, we observe a very weak correlation with r compared to $\mathcal{D}_t$ and $\mathcal{D}_a$.

To further investigate the reasons behind these correlation changes, we analyze the distribution of F1 scores, as shown in the distribution plot (see Figure 2). Following Finetuning, a significant proportion of models exhibit improved performance, particularly for easy and ambiguous instances. Interestingly, for hard instances, there is minimal difference in the distribution of F1 scores between the Evaluate and Finetune strategies. Specifically for dataset $\mathcal{D}_a$, performance declined after Finetuning, indicating that Finetuning on hard instances did not necessarily enhance model performance.

[17] We also experimented with other models such as XLNet (Yang et al. 2020), BERT (Devlin et al. 2019), and ALBERT (Lan et al. 2020) to categorize instances into different difficulty categories. However, as the average overlap percentage of instances between these models in their respective difficulty categories is higher, we decided to stick with RoBERTa.

[18] https://huggingface.co/oliverguhr/german-sentiment-bert

[19] Rank assigned by sorting F1-scores of models in the ascending order, 1: high-ranked, 500: least-ranked

[20] https://huggingface.co/ProsusAI/finbert

[21] https://huggingface.co/cardiffnlp/xlm-v-base-tweet-sentiment-ar

Despite being perceived as top-performing models as evident from their popularity, these high-ranked models within M_{500} did not consistently outperform other low-ranked (or less popular) models when handling instances of varying difficulty or from different domains.

> **Takeaway:** Contrary to common belief, high popularity does not always guarantee superior model performance.

4.2 Manual Inspection of M_{500}

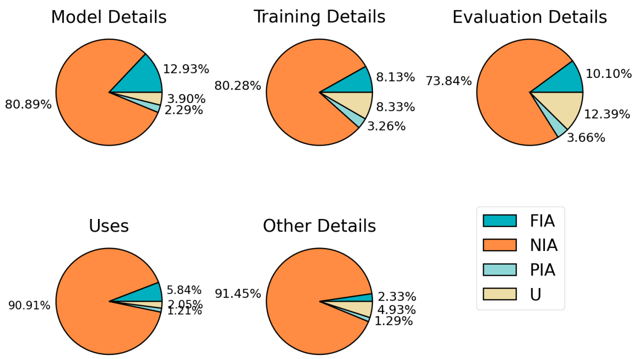

Figure 3: Results of manual inspection of model cards as described in §3.1. Each pie chart corresponds to a specific section within the model card.

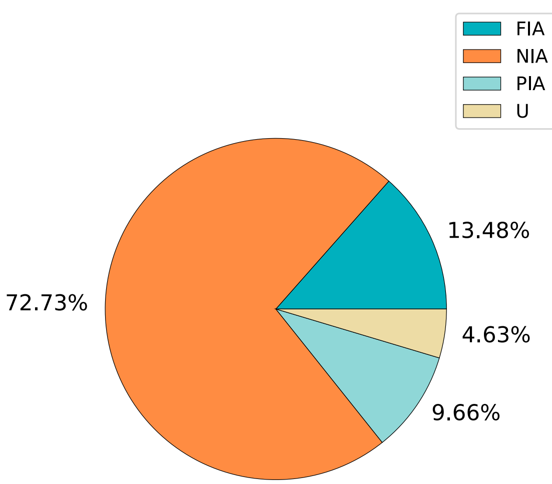

Figure 4: Results of manual inspection of class labels for the M_{500} models.

As detailed in §3.1, post-inspection of M_{500}, depicted in Figure 3, a substantial lack of information emerges. Approximately 81% of models lack all the details in the **Model Details** section of the model card, while 80% do not provide insights into specific training processes and corresponding details. Additionally, about 74% did not provide any information about model evaluation (**Evaluation Details**). Notably, 91% of models lack complete information concerning **Uses** and **Other Details**.

A more granular analysis of Model Card sections concerning Figure 5, focusing on mandatory (non-optional) elements per section (see Figure 1), reveals significant gaps. For **Training Details**, nearly 68% of models provide no information about the training data, approximately 85% omit details about the training process, and about 59% fail to disclose hyperparameter information. Similar patterns are observed in **Evaluation** and **Model Details**. Within the **Uses** section, essential details such as **Direct Use** (85%), **Out-of-Scope Use** (99%), **Bias Risks, and Limitations** (96%), and **Recommendations** (96%) are predominantly absent. The absence of these critical details raises questions about the models' suitability for downstream tasks.

During the inspection, we observe the absence of label mapping, particularly for categorical integer labels. This absence hinders the interpretation of which categorical label corresponds to which sentiments (Positive, Negative, or Neutral). In cases where no label information is present, we perform queries using HF Inference API[22] on their model page, categorizing these models under PIA. In instances where label outputs are ambiguous, such as `Label_0`, `Label_1`, and `Label_2`, it becomes challenging to determine the corresponding sentiment. These issues are categorized as NIA and are observed in 23% of models, as illustrated in Figure 4.

Does the Performance of a Model Correlate With How Well Documented Its Model Cards Are? We conducted a qualitative study of the M_{500} model cards through manual inspection as detailed in §3.2. Our investigation reveals a weak positive correlation between the comprehensiveness of the information provided in the model cards and the performance of the respective models, as evident from Table 4. This suggests that as the amount of full information available in the model cards (FIA) increases, there is a slight tendency for models to perform better on easy and ambiguous instances as compared to hard instances. However, the correlation is weak, so this relationship is not strong or highly predictive. This trend remains consistent across all categories of instances (easy, ambiguous, and hard) in dataset $\mathcal{D}$, with one exception: for hard instances of $\mathcal{D}_t$, the correlation score is statistically insignificant.

Interestingly, while some well-documented models perform well, others with similarly comprehensive documentation might not. Conversely, less-documented models occasionally outperform their well-documented counterparts, further weakening the overall correlation. This variability highlights that documentation alone is not a reliable indicator of model performance.

Overall, our findings indicate that models with more available information (FIA) did not consistently perform better in downstream tasks. Conversely, models with insufficient information (NIA) show a very weak negative correlation, suggesting that while these models tend to perform slightly

[22] https://huggingface.co/docs/api-inference/index

	$\mathcal{D}_t$			$\mathcal{D}_a$			$\mathcal{D}_f$	**r**		
	Easy	**Ambi**	**Hard**	**Easy**	**Ambi**	**Hard**	**Easy**	r_d	r_l	r_{da}
FIA	0.231*	0.221*	0.023	0.192*	0.196*	0.169*	0.151*	0.619*	0.527*	-0.081
PIA	0.112*	0.085	-0.071	0.067	0.076	0.081	0.066	0.508*	0.458*	-0.158*
NIA	-0.180*	-0.168*	-0.027	-0.155*	-0.166*	-0.145*	-0.165*	-0.591*	-0.491*	0.067

Table 4: Correlation values computed between the percentage of FIA/PIA/NIA (across all sections of the model card) and model performance (F1). Additionally, it includes correlation values calculated between the percentage of FIA/PIA/NIA and r (model popularity). Values highlighted with * represent statistical significant values with p-value < 0.05. For $\mathcal{D}_f$, we did not perform any evaluation strategy for ambiguous and hard instances, since the #instances categorized as ambiguous and hard is very low, which is insufficient for training and evaluation, therefore values are not calculated and displayed.

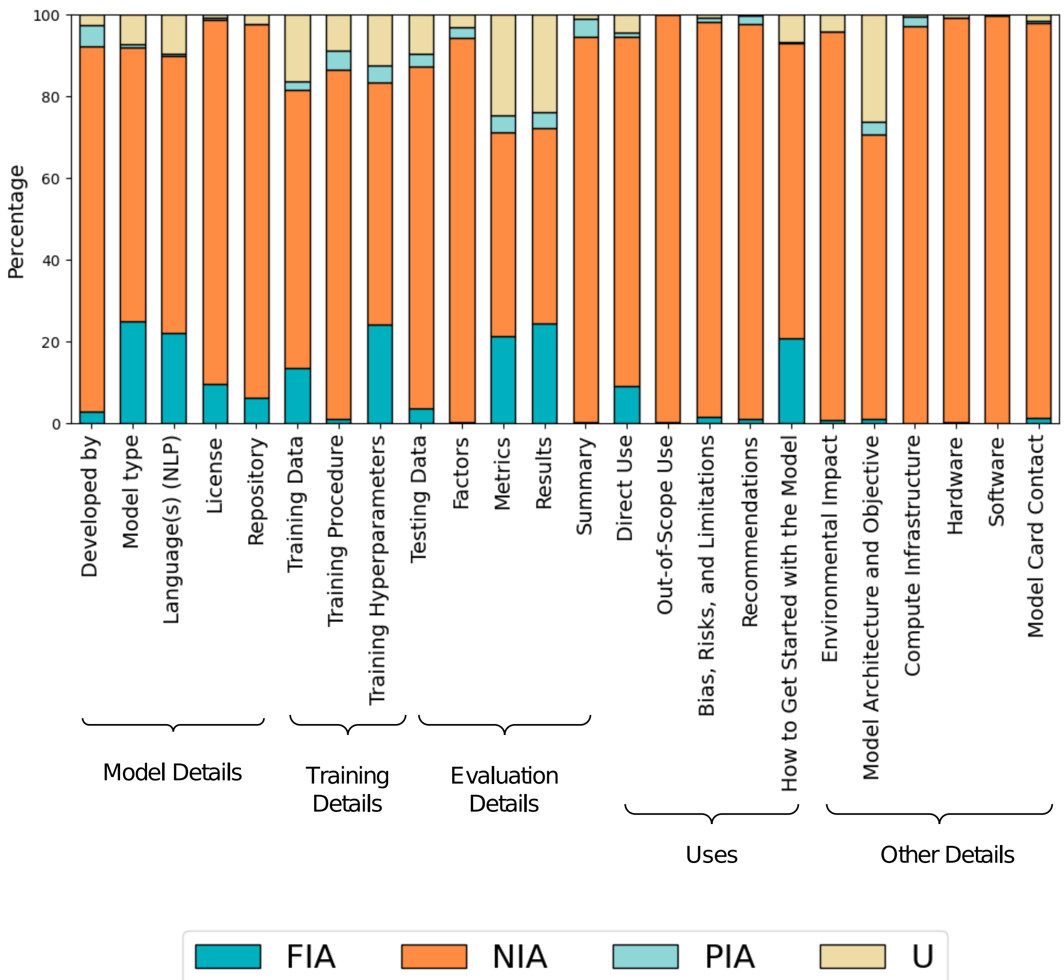

Figure 5: Results of manual inspection of model cards (granular analysis), as detailed in §3.1. Each bar corresponds to a specific element within a section of the model card.

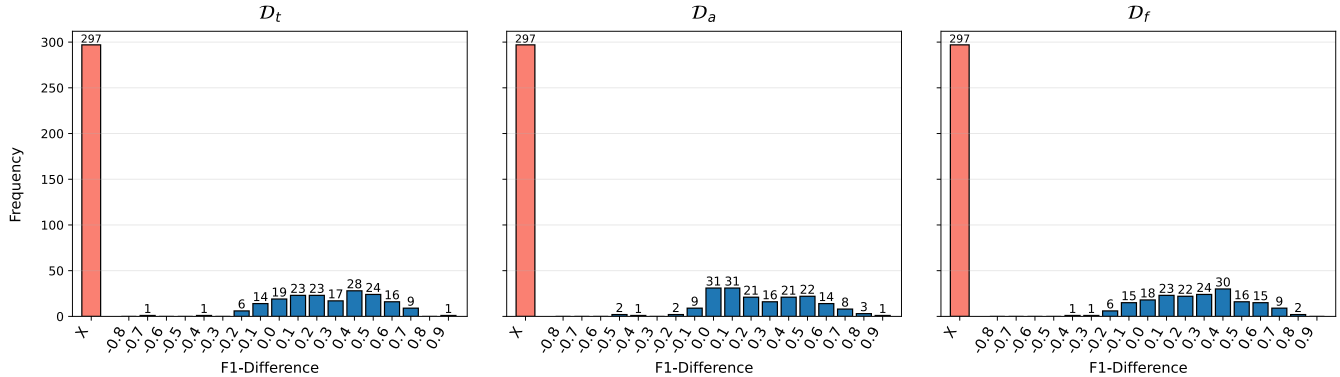

Figure 6: Difference between reported performance (F1) and empirical performance (F1) scores across datasets in $\mathcal{D}$, with F1-Difference on the X-axis and Frequency on y-axis. Empirical performance refers to the performance of M_{500} models evaluated on the full dataset of $\mathcal{D}$, while reported performance indicates the performance reported by the authors in their model cards. Cross (X) mark on the X-axis, with red colored bar represent the #models whose scores are not reported. For the mapping of model IDs to model names, refer Table 7 to Table 11 in the §A.

worse than those with more information, the weak correlation implies that this is not always the case. This trend is consistent across all categories in dataset $\mathcal{D}$, except for hard instances of $\mathcal{D}_t$ for which the correlation score is statistically insignificant. We did not draw any conclusions from PIA scores due to their statistical insignificance.

Takeaway: Comprehensive documentation does not necessarily gurantee better model performance.

Are Popular Models Well Documented? We continue our analysis by examining the correlation between the amount of information provided in model cards and r. Our analysis reveals a notable correlation between the comprehensiveness of information in model cards and their model rankings, as demonstrated in Table 4. Specifically, models with more information available (FIA/PIA) in their model cards tend to receive higher rankings in terms of r. Conversely, for models with insufficient information (NIA), the correlation is negative, indicating lower rankings (as r_d increases [eg., from r_d:1 to r_d:5], completeness rank (r_c)[23]decreases (eg., from r_c:5 to r_c:1)). This suggests as amount of information in model card decreases, there is a tendency for models to be less popular. These trends are consistent across all datasets. These observations underscore the importance of providing comprehensive information in model cards, as it appears to be correlated with the perceived trustworthiness and popularity of models for downstream tasks. It is important to note that this correlation does not imply causation. The relationship could be bidirectional, where more comprehensive documentation might enhance model popularity, and higher popularity could lead maintainers to invest more time in documenting the model. Further research is needed to explore this potential bidirectional relationship.

Takeaway: Models accompanied by more comprehensive details in their model cards tended to have higher popularity among users and vice-versa.

Does the Reported Performance Accurately Reflect Its True Capabilities? We further investigate whether the performance scores reported in the model cards of M_{500} accurately reflect the models' true performance.

As almost all of the models have not specified details about their evaluation datasets, we are unable to assess their performance directly on those datasets. However, as a proxy, we evaluate them on similar domains. For example, Model X is evaluated on Dataset Y (unknown), and we record its performance (as mentioned in the model card). We then compare this by evaluating Model X on a benchmark dataset from the same domain as Dataset Y.

To explore this further, we examine the performance of M_{500} on the complete dataset (combining train, validation, and test sets) of $\mathcal{D}$. Initially, we assess what percentage of models reported their scores on HF model cards. Our findings reveal that approximately 60% of models failed to report their scores[24]. Among those who did report their scores, we assess the validity of the reported scores. To accomplish this, we conduct evaluations of M_{500} models on the complete datasets of $\mathcal{D}$. Following the evaluations, we observe that a notable proportion ($\approx$ 88%) of models performed worse on $\mathcal{D}$ compared to their original performance on some other dataset (it may or may not be $\mathcal{D}$) as depicted in Figure 6. These findings suggest that many model authors do not thoroughly evaluate their models across diverse dataset domains. Even if evaluations are conducted, we recommend

[23]Completeness rank (r_c) is calculated by sorting percentage of FIA/PIA/NIA in descending order and assigning the ranks in ascending order, from 1 to 500

[24]We include models that have reported the macro F1 score for any given dataset and exclude the rest, which amounts to 10% of models. Additionally, we exclude approximately 4% of models from our analysis due to errors during model evluations.

that model authors accurately document the downstream usage of their models, their potential performance across various dataset domains, and any associated limitations.

> **Takeaway:** A staggering 88% of model authors overstated scores higher than what is reported in the model cards.

4.3 A Case Study on the Reddit Dataset

In the preceding sections, we examined the performance of M_{500} on popular benchmark datasets. Here, we further assess M_{500}'s capabilities by evaluating its performance on a new, unseen dataset. Specifically, we conduct a case study using a real-world dataset sourced from social media posts on Reddit[25].

Creation of Reddit Dataset We utilized the PRAW[26] library to scrape comments from the top 100 posts (of all time) of ten subreddits: Afghanistan (r/afghanistan), China (r/china), Bhutan (r/bhutan), Bangladesh (r/bangladesh), India (r/india), Sri Lanka (r/srilanka), Pakistan (r/pakistan), Maldives (r/maldives), Myanmar (r/myanmar), and Nepal (r/nepal). These subreddits primarily feature political content and this process gathered a total of 16,000 comments, including nested ones. After merging the comments, those containing URLs were excluded from subsequent sampling. Our focus narrowed to comments with a token count (excluding stopwords) ranging from 5 to 50 to maintain the quality of the text. We employed three SA models–cardiffnlp/twitter-roberta-base-sentiment-latest[27], finiteautomata/bertweet-base-sentiment-analysis[28], and Seethal/sentiment_analysis_generic_dataset[29] chosen for their popularity[30] to determine the majority sentiment for each comment. To ensure balanced representation, we randomly sampled 1,000 comments, aiming for an equal distribution of positive, negative, and neutral majority sentiments. The average sentence length for this dataset is 21 (words).

We then recruit four undergraduate students proficient in English and familiar with SA tasks to annotate the 1,000 instances. Before annotation, all annotators receive detailed instructions and examples of pre-annotated instances. Inter-annotator agreement is assessed using Krippendorff's α, yielding a score of 0.46.

Evaluation and Analysis Next, we evaluate the performance of the M_{500} on the 1,000-instance Reddit dataset and calculate the correlation between r and the M_{500} performance scores. We observe a very weak correlation between r and M_{500} performance, with correlation scores of 0.266, 0.136, and -0.1 for r_d, r_l, and r_{da} respectively, all of which are statistically significant ($p < 0.05$). We also report the mean and standard deviation of the reported F1 scores: (0.74, 0.17), compared to the mean and standard deviation of empirical F1 scores on the Reddit dataset: (0.41, 0.16) which indeed indicate poor generalization.

Dataset	Avg.	Min	Max
Reddit	20.8	6	50
$\mathcal{D}_t$	22.3	1	88
$\mathcal{D}_a$	42	3	808
$\mathcal{D}_f$	23	2	81

Table 5: The table displays average, min. and max. sentence length (#words) in the datasets studied, an insight into the data distribution.

The weak generalization observed may be attributed to a domain mismatch, as the Reddit dataset likely differs substantially in style, tone, and structure from the datasets on which the models were originally trained or fine-tuned. Training datasets are often curated from formal or task-specific texts, whereas Reddit's informal, conversational, and heterogeneous content poses distinct challenges that these models may not be equipped to address effectively. Furthermore, a potential popularity bias could influence model rankings, with higher-ranked models being favored for factors unrelated to their actual performance, such as ease of integration, familiarity within the community, or promotional efforts by model card authors. This bias may contribute to an inflated perception of their generalization capabilities.

Furthermore, the nature of Reddit comments, often characterized by informal language, slang, sarcasm, and context-dependent sentiment adds an additional layer of complexity. These features may be harder for models to interpret accurately, especially if they have not been exposed to similar linguistic characteristics during training.

Overall, these findings indicate that, although considered the top performing models due to their popularity, the M_{500} models demonstrate weak correlations and poor generalization on this dataset. This highlights the importance of fine-tuning on diverse and representative datasets to improve generalization across varying domains and dataset difficulty types.

> **Takeaway:** Although perceived as top models, M_{500} displays poor generalization performance.

4.4 Key Recommendations

This section provides empirically-backed recommendations for two key user groups: those downloading pre-trained models for downstream tasks and those contributing models to public repositories. While some of these recommendations might be intuitively known, our study emphasizes their importance through empirical evidence.

[26]https://praw.readthedocs.io/en/stable/

[27]https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[28]https://huggingface.co/finiteautomata/bertweet-base-sentiment-analysis

[29]https://huggingface.co/Seethal/sentiment_analysis_generic_dataset

[30]As of August 20th, 2023

[30]https://www.reddit.com/

For users downloading models:

- *Don't rely solely on popularity:* Avoid selecting models based solely on their high number of downloads or likes. Popularity does not always correlate with performance quality. For instance, `oliverguhr/german-sentiment-bert`[31] (r_d: 12, F1-rank: 266 on $\mathcal{D}_t$ (easy)) demonstrates poor generalization despite its high popularity (refer to §4.1).
- *Check for comprehensive documentation:* While detailed and well-structured documentation can aid in understanding a model's usage, do not assume that well-documented models always perform better. For example, `HerbertAIHug/Finetuned-Roberta-Base-Sentiment-identifier`[32] has detailed documentation but does not consistently outperform less-documented models (refer to §4.2). Ideal documentation examples include `distilbert/distilgpt2`[33].
- *Evaluate model suitability:* Consider the model's documented performance on relevant benchmarks or similar tasks to ensure its suitability for a specific use case (refer to §4.2). For instance, `cross-encoder/ms-marco-MiniLM-L-6-v2`[34] and `sismetanin/sbert-ru-sentiment-krnd`[35] demonstrate strong task-specific performance.

For users contributing models to public repositories:

- *Emphasize thorough documentation:* Comprehensive documentation is critical to making a model popular, trusted, and widely accepted on platforms like HF. Include details on training data, procedures, hyperparameters, and any limitations or biases (refer to §4.2). Examples of well-documented models include `cardiffnlp/twitter-roberta-base-sentiment-latest`[36].
- *Highlight performance metrics:* Provide detailed performance metrics and benchmarks to help users understand the model's strengths and weaknesses. Transparency in performance builds credibility and trust (refer to §4.2). For example, `oliverguhr/german-sentiment-bert`[37] and `sismetanin/sbert-ru-sentiment-krnd`[38] explicitly document their performance metrics.
- *Engage with the community:* Actively respond to user feedback and queries. Community engagement improves the model's reputation and can highlight areas for improvement. For example, `google-bert/bert-base-uncased`[39] maintains an active discussion platform.

[31] https://huggingface.co/oliverguhr/german-sentiment-bert

[32] https://huggingface.co/HerbertAIHug/Finetuned-Roberta-Base-Sentiment-identifier

[33] https://huggingface.co/distilbert/distilgpt2

[34] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

[35] https://huggingface.co/sismetanin/sbert-ru-sentiment-krnd

[36] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[37] https://huggingface.co/oliverguhr/german-sentiment-bert

[38] https://huggingface.co/sismetanin/sbert-ru-sentiment-krnd

5 Conclusions

In our study, we evaluated popular SA models on HF, exploring their performance across various datasets and difficulty categories to identify any correlation between popularity and performance. Surprisingly, we found no strong connection between popularity and overall model performance, with models lacking detailed information in their cards being less popular and performing worse. Our findings highlight the importance of thorough documentation and rigorous evaluation processes in NLP, emphasizing that popularity metrics alone are insufficient for assessing a model's suitability. Instead, we must develop more robust metrics that prioritize transparency, reproducibility, and comprehensive evaluation to drive advancements in NLP research and ensure the reliability of models in real-world applications.

Limitations

The current study is constrained to sentiment analysis tasks, limiting the generalizability of the findings to other domains within NLP. Similarly, the scope of the study is confined to three datasets, which may restrict the broader applicability of the results and observations. Additionally, resource limitations pose a significant bottleneck, preventing experiments on a very large number of models, large-sized (like Llama models) with respect to the #parameters, or across extensive text corpora.

Ethics Statement

Before participating in the evaluation, all human participants were provided with clear and comprehensive information about the nature and objectives of the study. Explicit informed consent was obtained from each participant before their involvement in the research.

We also recognize that subjective judgments, particularly in assessing model card completeness, may introduce potential biases in our evaluation. While we followed predefined guidelines and incorporated multiple annotators to improve consistency, some level of subjectivity remains. We consider fairness and transparency in our methodology a priority and encourage future work to explore more standardized evaluation frameworks.

Acknowledgements

We express our sincere gratitude to the University Grants Commission (UGC), Ministry of Education, Government of India, for funding the lead author's PhD through the UGC-JRF program, which made this work possible. We also thank the anonymous reviewers for their valuable feedback, which has greatly contributed to improving this paper.

[39] https://huggingface.co/google-bert/bert-base-uncased/discussions

References

Barabási, A.-L.; and Albert, R. 1999. Emergence of Scaling in Random Networks. *Science*, 286(5439): 509–512.

Barbieri, F.; Camacho-Collados, J.; Espinosa Anke, L.; and Neves, L. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. Online: Association for Computational Linguistics.

Castaño, J.; Mart'inez-Fern'andez, S.; Franch, X.; and Bogner, J. 2023. Analyzing the Evolution and Maintenance of ML Models on Hugging Face. *ArXiv*, abs/2311.13380.

Castaño, J.; Martínez-Fernández, S.; Franch, X.; and Bogner, J. 2023. Exploring the Carbon Footprint of Hugging Face's ML Models: A Repository Mining Study. In *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 1–12.

Crisan, A.; Drouhard, M.; Vig, J.; and Rajani, N. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 427–439. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2025-01-15.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Hussain, A.; Holla, N.; Mishra, P.; Yannakoudakis, H.; and Shutova, E. 2021. Towards a robust experimental framework and benchmark for lifelong language learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Jiang, W.; Synovic, N.; Hyatt, M.; Schorlemmer, T. R.; Sethi, R.; Lu, Y.-H.; Thiruvathukal, G. K.; and Davis, J. C. 2023. An Empirical Study of Pre-Trained Model Reuse in the Hugging Face Deep Learning Model Registry. In *Proceedings of the 45th International Conference on Software Engineering*, ICSE '23, 2463–2475. IEEE Press. ISBN 9781665457019.

Kadasi, P.; and Singh, M. 2023. Unveiling the Multi-Annotation Process: Examining the Influence of Annotation Quantity and Instance Difficulty on Model Performance. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1371–1388. Singapore: Association for Computational Linguistics.

Kathikar, A.; Nair, A.; Lazarine, B.; Sachdeva, A.; and Samtani, S. 2023. Assessing the Vulnerabilities of the Open-Source Artificial Intelligence (AI) Landscape: A Large-Scale Analysis of the Hugging Face Platform. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 1–6.

Keung, P.; Lu, Y.; Szarvas, G.; and Smith, N. A. 2020. The Multilingual Amazon Reviews Corpus. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4563–4568. Online: Association for Computational Linguistics.

Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; and Shah, M. 2022. Transformers in Vision: A Survey. *ACM Comput. Surv.*, 54(10s).

Kurihara, K.; Kawahara, D.; and Shibata, T. 2022. JGLUE: Japanese General Language Understanding Evaluation. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2957–2966. Marseille, France: European Language Resources Association.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv:1909.11942.

Liu, J.; Li, W.; Jin, Z.; and Diab, M. 2024. Automatic Generation of Model and Data Cards: A Step Towards Responsible AI. arXiv:2405.06258.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Malo, P.; Sinha, A.; Korhonen, P.; Wallenius, J.; and Takala, P. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4): 782–796.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 220–229. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Pepe, F.; Nardone, V.; Mastropaolo, A.; Bavota, G.; Canfora, G.; and Di Penta, M. 2024. How do Hugging Face Models Document Datasets, Bias, and Licenses? An Empirical Study. In *Proceedings of the 32nd IEEE/ACM International Conference on Program Comprehension*, ICPC '24, 370–381. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705861.

Shah, R.; Chawla, K.; Eidnani, D.; Shah, A.; Du, W.; Chava, S.; Raman, N.; Smiley, C.; Chen, J.; and Yang, D. 2022.

When FLUE Meets FLANG: Benchmarks and Large Pre-trained Language Model for Financial Domain. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2322–2335. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Singh, S.; Lodwal, H.; Malwat, H.; Thakur, R.; and Singh, M. 2023. Unlocking Model Insights: A Dataset for Automated Model Card Generation. *ArXiv*, abs/2309.12616.

Sipio, C. D.; Rubei, R.; Rocco, J. D.; Ruscio, D. D.; and Nguyen, P. T. 2024. Automated categorization of pre-trained models for software engineering: A case study with a Hugging Face dataset. arXiv:2405.13185.

Spearman, C. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1): 72–101.

Swayamdipta, S.; Schwartz, R.; Lourie, N.; Wang, Y.; Hajishirzi, H.; Smith, N. A.; and Choi, Y. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9275–9293. Online: Association for Computational Linguistics.

Taraghi, M.; Dorcelus, G.; Foundjem, A. T.; Tambon, F.; and Khomh, F. 2024. Deep Learning Model Reuse in the HuggingFace Community: Challenges, Benefit and Trends. *ArXiv*, abs/2401.13177.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, N.; Yang, H.; and Wang, C. 2023. FinGPT: Instruction Tuning Benchmark for Open-Source Large Language Models in Financial Datasets. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In Liu, Q.; and Schlangen, D., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237.

Zhang, W.; Deng, Y.; Liu, B.; Pan, S. J.; and Bing, L. 2023. Sentiment Analysis in the Era of Large Language Models: A Reality Check. arXiv:2305.15005.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? N/A
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? N/A
   (g) Did you discuss any potential misuse of your work? N/A
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes,in the Abstract

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, Refer Table 1

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, in Hyperparameters section. refer Appendix C

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? N/A

(c) Did you include any new assets in the supplemental material or as a URL? No

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, but all the data and assets have been taken from the Open-Source platforms like Hugging Face

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? N/A

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? N/A

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? N/A

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, Refer the Instruction section in Appendix E

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? No, all the participants are UG students who were willing to volunteer

(d) Did you discuss how data is stored, shared, and deidentified? N/A

A Appendix

A.1 More details on Models Filtering

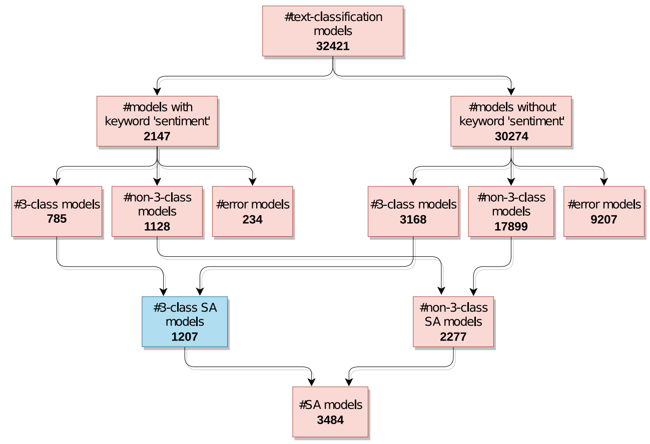

Figure 7: Models Filtering

Dataset	Learning Rate	Per Device Train Batch Size	#Train Epochs
$\mathcal{D}_t$	1.9298e-05	128	5
$\mathcal{D}_f$	1.1470e-05	16	19
$\mathcal{D}_a$	6.2177e-06	32	3

Table 6: Optimal Hyperparameters

In filtering the models, we initially focused on those categorized under the "Text Classification" task on HF. These models are categorized into two groups: those containing the keyword "sentiment" and those without it.To distinguish between these groups of models of 3-class and non-3-class, we ran an example input text through each model in the respective groups and analyzed the number of output classes. This process result in four distinct groups (3-class-with-sentiment, 3-class-without-sentiment, non-3-class-with-sentiment, non-3-class-without-sentiment).

For models lacking the "sentiment" keyword in their names. we conducted additional checks to determine whether they qualified as SA models. Specifically, we examine their output classes to identify labels such as "positive," "very positive," "neutral," "negative," "very negative," and variations like"pos," "neg," "neu," "positiv," "notr," "negativ," or uppercase equivalents (e.g., "POSITIVE," "NEGATIVE"). Substring matching is used to include variations of these labels. Models producing such outputs are classified as SA models. Conversely, models with the "sentiment" keyword are directly classified as SA models. We restrict our analysis to 3-class models to maintain simplicity and avoid complexity introduced by models with more than three class labels. This process yielded a total of 1207 3-class SA models as depicted in Figure 7, from which the M_{500} is selected based on r_d.

Model ID	Model Name	Model ID	Model Name
M1	cardiffnlp/twitter-roberta-base-sentiment-latest	M51	KBLab/robust-swedish-sentiment-multiclass
M2	mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis	M52	amansolanki/autonlp-Tweet-Sentiment-Extraction-20114061
M3	cardiffnlp/twitter-roberta-base-sentiment	M53	knkarthick/Sentiment-Analysis
M4	lxyuan/distilbert-base-multilingual-cased-sentiments-student	M54	Ammar-alhaj-ali/arabic-MARBERT-sentiment
M5	ProsusAI/finbert	M55	RogerKam/roberta_RCADE_fine_tuned_sentiment_covid_news
M6	yiyanghkust/finbert-tone	M56	Jean-Baptiste/roberta-large-financial-news-sentiment-en
M7	avichr/heBERT_sentiment_analysis	M57	FinanceInc/auditor_sentiment_finetuned
M8	finiteautomata/bertweet-base-sentiment-analysis	M58	akoksal/bounti
M9	cardiffnlp/twitter-xlm-roberta-base-sentiment	M59	MarieAngeA13/Sentiment-Analysis-BERT
M10	finiteautomata/beto-sentiment-analysis	M60	Venkatesh4342/distilbert-helpdesk-sentiment
M11	Seethal/sentiment_analysis_generic_dataset	M61	Souvikcmsa/BERT_sentiment_analysis
M12	oliverguhr/german-sentiment-bert	M62	CAMeL-Lab/bert-base-arabic-camelbert-ca-sentiment
M13	CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment	M63	ka05ar/banglabert-sentiment
M14	ahmedrachid/FinancialBERT-Sentiment-Analysis	M64	cardiffnlp/xlm-twitter-politics-sentiment
M15	cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual	M65	alexandrainst/da-sentiment-base
M16	koheiduck/bert-japanese-finetuned-sentiment	M66	RecordedFuture/Swedish-Sentiment-Fear
M17	blanchefort/rubert-base-cased-sentiment	M67	JP040/bert-german-sentiment-twitter
M18	MonoHime/rubert-base-cased-sentiment-new	M68	philschmid/distilbert-base-multilingual-cased-sentiment
M19	rohanrajpal/bert-base-multilingual-codemixed-cased-sentiment	M69	tkurtulus/TurkishAirlines-SentimentAnalysisModel
M20	snunlp/KR-FinBert-SC	M70	zainalq7/autotrain-NLU_crypto_sentiment_analysis-754123133
M21	cointegrated/rubert-tiny-sentiment-balanced	M71	sahri/indonesiasentiment
M22	sbcBI/sentiment_analysis_model	M72	yangheng/deberta-v3-large-absa-v1.1
M23	soleimanian/financial-roberta-large-sentiment	M73	Tejas3/distillbert_base_uncased_80_equal
M24	ayameRushia/bert-base-indonesian-1.5G-sentiment-analysis-smsa	M74	turing-usp/FinBertPTBR
M25	ayameRushia/roberta-base-indonesian-1.5G-sentiment-analysis-smsa	M75	mr4/phobert-base-vi-sentiment-analysis
M26	citizenlab/twitter-xlm-roberta-base-sentiment-finetunned	M76	elozano/tweet_sentiment_eval
M27	yangheng/deberta-v3-base-absa-v1.1	M77	cardiffnlp/bertweet-base-sentiment
M28	philschmid/distilbert-base-multilingual-cased-sentiment-2	M78	hakonmh/sentiment-xdistil-uncased
M29	climatebert/distilroberta-base-climate-sentiment	M79	FinScience/FS-distilroberta-fine-tuned
M30	j-hartmann/sentiment-roberta-large-english-3-classes	M80	rohanrajpal/bert-base-codemixed-uncased-sentiment
M31	Sigma/financial-sentiment-analysis	M81	seara/rubert-base-cased-russian-sentiment
M32	blanchefort/rubert-base-cased-sentiment-rusentiment	M82	cardiffnlp/xlm-roberta-base-sentiment-multilingual
M33	w11wo/indonesian-roberta-base-sentiment-classifier	M83	cardiffnlp/camembert-base-tweet-sentiment-fr
M34	mdhugol/indonesia-bert-sentiment-classification	M84	LondonStory/txlm-roberta-hindi-sentiment
M35	HooshvareLab/bert-fa-base-uncased-sentiment-digikala	M85	shashanksrinath/News_Sentiment_Analysis
M36	christian-phu/bert-finetuned-japanese-sentiment	M86	xaqren/sentiment_analysis
M37	lucas-leme/FinBERT-PT-BR	M87	emre/turkish-sentiment-analysis
M38	CAMeL-Lab/bert-base-arabic-camelbert-mix-sentiment	M88	RogerKam/roberta_fine_tuned_sentiment_newsmtsc
M39	pin/senda	M89	RecordedFuture/Swedish-Sentiment-Violence
M40	sbcBI/sentiment_analysis	M90	deepset/bert-base-german-cased-sentiment-Germeval17
M41	neuraly/bert-base-italian-cased-sentiment	M91	cardiffnlp/xlm-roberta-base-tweet-sentiment-pt
M42	hw2942/bert-base-chinese-finetuning-financial-news-sentiment-v2	M92	KernAI/stock-news-distilbert
M43	poom-sci/WangchanBERTa-finetuned-sentiment	M93	Timothy1337/finetuning-sentiment-all_df
M44	fergusq/finbert-finnsentiment	M94	blanchefort/rubert-base-cased-sentiment-rurewiews
M45	CouchCat/ma_sa_v7_distil	M95	Voicelab/herbert-base-cased-sentiment
M46	peerapongch/baikal-sentiment-ball	M96	svalabs/twitter-xlm-roberta-bitcoin-sentiment
M47	samayash/finetuning-financial-news-sentiment	M97	CAMeL-Lab/bert-base-arabic-camelbert-msa-sentiment
M48	seara/rubert-tiny2-russian-sentiment	M98	bardsai/finance-sentiment-de-base
M49	aari1995/German_Sentiment	M99	kinit/slovakbert-sentiment-twitter
M50	mdraw/german-news-sentiment-bert	M100	hanifnoerr/Fine-tuned-Indonesian-Sentiment-Classifier

Table 7: Split-1 of M_{500} Models

Model ID	Model Name	Model ID	Model Name
M101	ganeshkharad/gk-hinglish-sentiment	M151	eevvgg/StanceBERTa
M102	Kyle1668/boss-sentiment-bert-base-uncased	M152	mrcaelumn/yelp_restaurant_review_sentiment_analysis
M103	oferweintraub/bert-base-finance-sentiment-noisy-search	M153	MMG/xlm-roberta-base-sa-spanish
M104	RashidNLP/Finance-Sentiment-Classification	M154	finiteautomata/beto-headlines-sentiment-analysis
M105	atowey01/hostel-reviews-sentiment-model	M155	KernAI/community-sentiment-bert
M106	bardsai/twitter-sentiment-pl-base	M156	sismetanin/rubert_conversational-ru-sentiment-rureviews
M107	clips/republic	M157	cardiffnlp/twitter-roberta-base-2021-124m-sentiment
M108	larskjeldgaard/senda	M158	Tobias/bert-base-german-cased_German_Hotel_sentiment
M109	bardsai/finance-sentiment-zh-base	M159	tyqiangz/indobert-lite-large-p2-smsa
M110	data354/camembert-fr-covid-tweet-sentiment-classification	M160	Venkatesh4342/distilbert-helpdesk-sentence-sentiment
M111	nimaafshar/parsbert-fa-sentiment-twitter	M161	seninoseno/rubert-base-cased-sentiment-study-feedbacks-solyanka
M112	dnzblgn/BART_Sentiment_Classification	M162	hazrulakmal/distilbert-optimised-finetuned-financial-sentiment
M113	cardiffnlp/roberta-base-sentiment	M163	Softechlb/Sent_analysis_CVs
M114	rohanrajpal/bert-base-en-es-codemix-cased	M164	Anthos23/FS-distilroberta-fine-tuned
M115	mstafam/fine-tuned-bert-financial-sentiment-analysis	M165	francisco-perez-sorrosal/distilbert-base-uncased-finetuned-with-spanish-tweets-clf
M116	sismetanin/xlm_roberta_large-ru-sentiment-rureviews	M166	sismetanin/rubert-ru-sentiment-rureviews
M117	Davlan/naija-twitter-sentiment-afriberta-large	M167	Tobias/bert-base-uncased_English_Hotel_sentiment
M118	blanchefort/rubert-base-cased-sentiment-med	M168	z-dickson/multilingual_sentiment_newspaper_headlines
M119	abnersampaio/sentiment	M169	techthiyanes/Bert_Bahasa_Sentiment
M120	cardiffnlp/roberta-base-tweet-sentiment-en	M170	RogerKam/roberta_fine_tuned_sentiment_sst3
M121	BVK97/Discord-NFT-Sentiment	M171	nikunjbjj/jd-resume-model
M122	alger-ia/dziribert_sentiment	M172	eevvgg/PaReS-sentimenTw-political-PL
M123	pysentimiento/roberta-es-sentiment	M173	IAyoub/finetuning-bert-sentiment-reviews-2
M124	l3cube-pune/MarathiSentiment	M174	Cristian-dcg/beto-sentiment-analysis-finetuned-onpremise
M125	m3hrdadfi/albert-fa-base-v2-sentiment-multi	M175	NYTK/sentiment-ohb3-hubert-hungarian
M126	bardsai/finance-sentiment-pl-fast	M176	cardiffnlp/xlm-v-base-tweet-sentiment-de
M127	Jorgeutd/bert-base-uncased-finetuned-surveyclassification	M177	warwickai/fin-perceiver
M128	nickwong64/bert-base-uncased-finance-sentiment	M178	mdeniz1/turkish-sentiment-analysis-bert-base-turkish-uncased
M129	ncduy/phobert-large-finetuned-vietnamese_students_feedback	M179	TFLai/turkish-bert-128k-sentiment
M130	Yah216/Sentiment_Analysis_CAMelBERT_msa_sixteenth_HARD	M180	benjaminbeilharz/bert-base-uncased-sentiment-classifier
M131	mwz/RomanUrduClassification	M181	RashidNLP/Amazon-Deberta-Base-Sentiment
M132	eevvgg/sentimenTw-political	M182	blinjrm/finsent
M133	yj2773/hinglish11k-sentiment-analysis	M183	Kapiche/twitter-roberta-base-sentiment-latest
M134	cardiffnlp/bert-base-multilingual-cased-sentiment-multilingual	M184	bardsai/finance-sentiment-zh-fast
M135	pysentimiento/roberta-targeted-sentiment-analysis	M185	EMBEDDIA/sloberta-tweetsentiment
M136	Souvikcmsa/SentimentAnalysisDistillBERT	M186	cassiepowell/RoBERTa-large-mnli-for-agreement
M137	candra/indobertweet-sentiment2	M187	oandreae/financial_sentiment_model
M138	sismetanin/sbert-ru-sentiment-rureviews	M188	cardiffnlp/xlm-roberta-base-tweet-sentiment-fr
M139	rohanrajpal/bert-base-en-hi-codemix-cased	M189	papepipopu/trading_ai
M140	sismetanin/sbert-ru-sentiment-krnd	M190	Elron/deberta-v3-large-sentiment
M141	DunnBC22/distilbert-base-uncased-Financial_Sentiment_Analysis	M191	meghanabhange/Hinglish-Bert-Class
M142	niksmer/RoBERTa-RILE	M192	EMBEDDIA/bertic-tweetsentiment
M143	l3cube-pune/marathi-sentiment-md	M193	Sonny4Sonnix/twitter-roberta-base-sentimental-analysis-of-covid-tweets
M144	ZiweiChen/FinBERT-FOMC	M194	EMBEDDIA/rubert-tweetsentiment
M145	RogerKam/roberta_fine_tuned_sentiment_financial_news	M195	arjuntheprogrammer/distilbert-base-multilingual-cased-sentiment-2
M146	akahana/indonesia-sentiment-roberta	M196	bowipawan/bert-sentimental
M147	amphora/KorFinASC-XLM-RoBERTa	M197	zabiullah/autotrain-customers_email_sentiment-3449294006
M148	abhishek/autonlp-swahili-sentiment-615517563	M198	slisowski/stock_sentiment_hp
M149	m3hrdadfi/albert-fa-base-v2-sentiment-digikala	M199	bardsai/finance-sentiment-pl-base
M150	classla/bcms-bertic-parlasent-bcs-ter	M200	adam-chell/tweet-sentiment-analyzer

Table 8: Split-2 of M_{500} Models

Model ID	Model Name	Model ID	Model Name
M201	Manauu17/enhanced_roberta_sentiments_es	M251	ymcnabb/dutch_threeway_sentiment_classification_v2
M202	lucaordronneau/twitter-roberta-base-sentiment-latest-finetuned-FG-SINGLE_SENTENCE-NEWS	M252	sudhanvasp/Sentiment-Analysis
M203	cardiffnlp/twitter-roberta-base-dec2021-sentiment	M253	hazrulakmal/bert-base-uncased-finetuned
M204	Splend1dchan/bert-base-uncased-slue-goldtrascription-e3-lr1e-4	M254	ramnika003/autotrain-sentiment_analysis_project-705021428
M205	everyl12/crisis_sentiment_roberta	M255	veb/twitch-roberta-base-sentiment-latest
M206	DunnBC22/bert-base-uncased-Twitter_Sentiment_Analysis_v2	M256	Abubakari/finetuned-Sentiment-classfication-ROBERTA-model
M207	Hyeonseo/ko-finance_news_classifier	M257	kullackaan/sentiment-tweets
M208	bardsai/twitter-sentiment-pl-fast	M258	chrommium/sbert_large-finetuned-sent_in_news_sents_3lab
M209	EMBEDDIA/english-tweetsentiment	M259	rasmodev/Covid-19_Sentiment_Analysis_RoBERTa_Model
M210	DSI/human-directed-sentiment	M260	Mawulom/Fine-Tuned-Bert_Base_Cased_Sentiment_Analysis
M211	hw2942/bert-base-chinese-finetuning-financial-news-sentiment	M261	ikoghoemmanuell/finetuned_sentiment_modell
M212	Tejas3/distillbert_base_uncased_80	M262	mtyrrell/CPU_Transport_GHG_Classifier
M213	bardsai/finance-sentiment-es-base	M263	mlkorra/obgv-gender-bert-hi-en
M214	sismetanin/rubert_conversational-ru-sentiment-sentirueval2016	M264	KarelDO/lstm.CEBaB_confounding.uniform.absa.5-class.seed_42
M215	arize-ai/distilbert_reviews_with_language_drift	M265	francisco-perez-sorrosal/dccuchile-distilbert-base-spanish-uncased-finetuned-with-spanish-tweets-clf
M216	Theivaprakasham/sentence-transformers-paraphrase-MiniLM-L6-v2-twitter_sentiment	M266	Hyeonseo/finance_news_classifier
M217	gabrielyang/finance_news_classifier-KR_v7	M267	LYTinn/finetuning-sentiment-model-tweet-gpt2
M218	researchaccount/sa_sub4	M268	francisco-perez-sorrosal/dccuchile-distilbert-base-spanish-uncased-finetuned-with-spanish-tweets-clf-cleaned-ds
M219	Abdelrahman-Rezk/emotion-english-distilroberta-base-fine_tuned_for_amazon_reviews_english_3	M269	cassiepowell/LaBSE-for-agreement
M220	TankuVie/bert-base-multilingual-uncased-vietnamese_sentiment_analysis	M270	francisco-perez-sorrosal/distilbert-base-multilingual-cased-finetuned-with-spanish-tweets-clf
M221	researchaccount/sa_sub1	M271	cardiffnlp/xlm-v-base-tweet-sentiment-en
M222	laurens88/finetuning-crypto-tweet-sentiment-test2	M272	francisco-perez-sorrosal/distilbert-base-multilingual-cased-finetuned-with-spanish-tweets-clf-cleaned-ds
M223	Narsil/finbert2	M273	dpeinado/twitter-roberta-base-sentiment-latest-apple-opinions
M224	Anthos23/my-awesome-model	M274	course5i/SEAD-L-6_H-384_A-12-mnli
M225	aXhyra/presentation_sentiment_1234567	M275	DunnBC22/distilbert-base-uncased-US_Airline_Twitter_Sentiment_Analysis
M226	hw2942/bert-base-chinese-finetuning-financial-news-sentiment-test	M276	Psunrise/finetuning-customer-sentiment-model-300-samples
M227	ScriptEdgeAI/MarathiSentiment-Bloom-560m	M277	AGudden/xlm-roberta-base-finetuned-marc
M228	Teeto/reviews-classification	M278	course5i/SEAD-L-6_H-256_A-8-mnli
M229	hilmansw/indobert-finetuned-sentiment-happiness-index	M279	eduardopds/distilbert-base-uncased-tweets
M230	Narsil/finbert-slow	M280	maclean-connor96/feedier-french-books
M231	BramVanroy/bert-base-multilingual-cased-hebban-reviews	M281	abnersampaio/sentimentv2
M232	cnut1648/biolinkbert-mnli	M282	CultureBERT/roberta-large-clan
M233	sismetanin/rubert_conversational-ru-sentiment-krnd	M283	CultureBERT/roberta-large-hierarchy
M234	Cloudy1225/stackoverflow-roberta-base-sentiment	M284	jayantapaul888/twitter-data-microsoft-deberta-base-mnli-sentiment-finetuned-memes
M235	poerwiyanto/bert-base-indonesian-522M-finetuned-sentiment	M285	francisco-perez-sorrosal/distilbert-base-uncased-finetuned-with-spanish-tweets-clf-cleaned-ds
M236	mirfan899/da_spacy_sentiment	M286	traptrip/rosatom_survey_sentiment_classifier
M237	GhylB/Sentiment_Analysis_DistilBERT	M287	tanoManzo/roberta-attitude
M238	Vasanth/bert-stock-sentiment-analyzer	M288	Christiansg/finetuning-sentiment_spanish-amazon-group23
M239	cardiffnlp/xlm-v-base-tweet-sentiment-pt	M289	sasha/autotrain-RobertaBaseTweetEval-1281048989
M240	sakasa007/finetuning-sentiment-text-mining	M290	IngeniousArtist/distilbert-finance
M241	CK42/sentiment_analysis_sbcBI	M291	Anthos23/FS-finbert-fine-tuned-f1
M242	kartashoffv/vashkontrol-sentiment-rubert	M292	cardiffnlp/xlm-roberta-base-tweet-sentiment-es
M243	dogruermikail/bert-fine-tuned-stock-sentiment-uncased	M293	cardiffnlp/xlm-roberta-base-tweet-sentiment-en
M244	researchaccount/sa_sub5	M294	hw2942/bert-base-chinese-finetuning-financial-news-sentiment-test1
M245	madmancity/bert2	M295	sasha/autotrain-RobertaBaseTweetEval-1281048990
M246	bright1/fine-tuned-twitter-Roberta-base-sentiment	M296	CultureBERT/roberta-large-adhocracy
M247	debashish68/roberta-sent-generali	M297	ruanchaves/bert-base-portuguese-cased-porsimplessent
M248	BramVanroy/xlm-roberta-base-hebban-reviews	M298	sasha/autotrain-RobertaBaseTweetEval-1281048986
M249	fassahat/anferico-bert-for-patents-finetuned-557k-patent-sentences	M299	bardsai/finance-sentiment-ja-base
M250	KarelDO/lstm.CEBaB_confounding.food_service_positive.absa.5-class.seed_42	M300	ruanchaves/bert-base-portuguese-cased-assin-entailment

Table 9: Split-3 of M_{500} Models

Model ID	Model Name	Model ID	Model Name
M301	ruanchaves/mdeberta-v3-base-assin-entailment	M351	deansaco/Roberta-base-financial-sentiment-analysis
M302	Nelver28/sentiment-analysis-positive-mixed-negative	M352	bvint/autotrain-sphere-lecture-demo-1671659193
M303	hazrulakmal/benchmark-finetuned-distilbert	M353	Svetlana0303/Classfication_RoBERTa
M304	gyesibiney/covid-tweet-sentimental-Analysis-roberta	M354	jayantapaul888/twitter-data-pysentimiento-robertuito-sentiment-finetuned-memes
M305	Rem59/autotrain-Test_2-789524315	M355	dshin/my_awesome_model
M306	ozoora/rubert-4.1poi	M356	ongknsro/ACARISBERT-DistilBERT
M307	profoz/deploy-mlops-demo	M357	Cincin-nvp/NusaX-senti_XLM-R
M308	CultureBERT/roberta-large-market	M358	tanoManzo/distilbert-attitude
M309	philschmid/finbert-tone-endpoint-ds	M359	slickdata/finetuned-Sentiment-classfication-ROBERTA-model
M310	ruanchaves/mdeberta-v3-base-porsimplessent	M360	vectorizer/sentiment_analysis_93k_entries
M311	ldeb/solved-finbert-tone	M361	Bennet1996/finetuning-ESG-sentiment-model-distilbert
M312	Venkatesh4342/bert-base-uncased-finetuned-fin	M362	adnanakbr/bert-base-multilingual-uncased-sentiment-fine_tuned_for_amazon_english_reviews_on_200K_review_v2
M313	cardiffnlp/xlm-roberta-base-tweet-sentiment-ar	M363	intanm/sa10-clm-20230403-001-3
M314	DingYao/autotrain-fbert-singlish-5-1943965533	M364	cardiffnlp/mbert-base-tweet-sentiment-fr
M315	sasha/autotrain-BERTBase-TweetEval-1281248998	M365	chinmayapani/panich
M316	MavisAJ/Sentiment_analysis_roberta_model	M366	davidchiii/news-headlines
M317	UchihaMadara/phobert-finetuned-sentiment-analysis	M367	Svetlana0303/Classfication_longformer
M318	cardiffnlp/mbert-base-tweet-sentiment-pt	M368	HerbertAIHug/Finetuned-Roberta-Base-Sentiment-identifier
M319	Venkatesh4342/xlm-roberta-helpdesk-sentiment	M369	cardiffnlp/xlm-v-base-tweet-sentiment-fr
M320	sasha/autotrain-BERTBase-TweetEval-1281248996	M370	Svetlana0303/Classfication_AlBERT
M321	DunnBC22/fnet-large-Financial_Sentiment_Analysis_v3	M371	gr8testgad-1/sentiment_analysis
M322	DingYao/autotrain-fbert-singlish-1755361190	M372	cardiffnlp/mbert-base-tweet-sentiment-ar
M323	IsaacSarps/sentiment_analysis	M373	ecabott/nepali-sentiment-analyzer
M324	sasha/autotrain-DistilBERT-TweetEval-1281148991	M374	cruiser/distilbert_model_kaggle_200_epoch
M325	cardiffnlp/xlm-v-base-tweet-sentiment-es	M375	nlp-chula/sentiment-finnlp-th
M326	cardiffnlp/xlm-roberta-base-tweet-sentiment-de	M376	giotvr/portuguese-nli-3-labels
M327	DingYao/autotrain-fbert-singlish-2-1937065404	M377	Pendo/finetuned-Sentiment-classfication-ROBERTA-Base-model
M328	dogruermikail/bert-fine-tuning-sentiment-stocks-analyis-uncased	M378	raygx/BERT-NepSA-domainAdapt
M329	l3cube-pune/marathi-sentiment-subtitles	M379	incredible45/News-Sentimental-model-Buy-Neutral-Sell
M330	ultraleow/cloud4bert	M380	cardiffnlp/xlm-v-base-tweet-sentiment-ar
M331	aarnphm/multi-length-text-classification-pipeline	M381	intanm/mlm_v1_20230327_fin_sa_10
M332	israel/testing_model	M382	RogerB/kin-sentiC
M333	raygx/BERT-NepSA-T2	M383	intanm/sa100-mlm-20230403-001-2
M334	devtanumisra/finetuning-sentiment-model-deberta-smote	M384	raoh/yiran-nlp4
M335	cardiffnlp/xlm-roberta-base-tweet-sentiment-it	M385	CMunch/fine_tuned_dota
M336	sairahul5223/autotrain-auto-train-intent-classification-20220928-1584756071	M386	asaderu-ai/ssclass_best
M337	sasha/autotrain-DistilBERT-TweetEval-1281148992	M387	salohnana2018/ABSA-single-domainAdapt-bert-base-MARBERT2-HARD
M338	cardiffnlp/mbert-base-tweet-sentiment-it	M388	cruiser/roberta_tweet_eval_finetuned
M339	antypasd/twitter-roberta-base-sentiment-earthquake	M389	sara-nabhani/ML-ns-bert-base-uncased
M340	QuophyDzifa/Sentiment-Analysis-Model	M390	raygx/distilBERT-NepSA
M341	fassahat/anferico-bert-for-patents-finetuned-150k-sentences	M391	Ausbel/Vaccine-tweet-sentiments-analysis-model-2
M342	Himanshusingh/finetunedfinbert-model	M392	salohnana2018/MARBERTV02ABSAnew
M343	SiddharthaM/twitter-data-bert-base-multilingual-uncased-hindi-only-memes	M393	duwuonline/mymodel-classify-sentiment
M344	cardiffnlp/xlm-v-base-tweet-sentiment-it	M394	AlonCohen/RuSentNE-test
M345	fassahat/distilbert-base-uncased-finetuned-557k-patent-sentences	M395	memotirre90/Equipo16_gpt2-HotelSentiment
M346	DunnBC22/fnet-base-Financial_Sentiment_Analysis	M396	neojex/testing
M347	brema76/vaccine_event_it	M397	xyu1163/Testmodel_sentiment_with_3_labels
M348	UholoDala/tweet_sentiments_analysis_bert	M398	jcy204/wind_model2
M349	W4nkel/microsoftTurkishTrain	M399	dhikaardianto/TiktokSentimentIndoBertTwitter
M350	aisyahhrazak/distilbert-mooc-review-sentiment	M400	midwinter73/dipterv6

Table 10: Split-4 of M_{500} Models

Model ID	Model Name	Model ID	Model Name
M401	salohnana2018/ARABERT02-best-trail	M451	yanezh/twiiter_try11_fold2
M402	cruiser/final_model	M452	Yt99/SFinBERT
M403	bekbote/autotrain-dl-phrasebank-53436126044	M453	yanezh/twiiter_try5_fold0
M404	senfu/bert-base-uncased-top-pruned-mnli	M454	abnersampaio/sentiment3
M405	abnersampaio/sentiment3.5	M455	yanezh/twiiter_try10_fold4
M406	intanm/fin-sa-post-trained-indobert-base-finreport	M456	intanm/fin-sa-post-trained-indobert-base-finnews-p2-001
M407	kreynolds03/Solomon	M457	yanezh/twiiter_try13_fold1
M408	rod16/v1_finetuning-sentiment-model-news-samples	M458	yanezh/twiiter_try4_fold2
M409	luissgtorres/Bert_sentiment_analysis_Indata	M459	langecod/Financial_Phrasebank_RoBERTa
M410	intanm/mlm_v1_20230327_fin_sa_70	M460	BeChi87/train_model_Lastest1
M411	Svetlana0303/Classfication_electra	M461	intanm/fin-sa-post-trained-indobert-base-combined
M412	intanm/mlm_v1_20230327_fin_sa_20	M462	evendivil/finetuning-sentiment-model-3000-samples
M413	Sonny4Sonnix/movie_sentiment_trainer	M463	yanezh/twiiter_try13_fold3
M414	A-Funakoshi/bert-finetuned-multilingual-sentiments	M464	gabrielkytz/finetuning-sentiment-model-3000-samples
M415	salohnana2018/ARABERT02	M465	ErisGrey/my_models
M416	intanm/mlm_v1_20230327_fin_sa_100	M466	HerbertAIHug/finetuned_sentiment_analysis_modell
M417	intanm/mlm_v1_20230327_fin_sa_90	M467	yanezh/twiiter_try5_fold3
M418	A-Funakoshi/bert-base-japanese-v3-wrime-sentiment	M468	Abdelrahman-Rezk/emotion-english-distilroberta-base-fine_tuned_for_amazon_english_reviews_V03_on_100K_review
M419	Q317/EmoraBert	M469	yanezh/twiiter_try13_fold4
M420	NewtonKimathi/Covid_Vaccine_Sentiment_Analysis_Roberta_Model	M470	yanezh/twiiter_try4_fold3
M421	yanezh/twiiter_try13_fold2	M471	yanezh/twiiter_try6_fold1
M422	quesmed/tone	M472	BeChi87/train_model_Lastestttt
M423	intanm/baseline_fin_sa_60	M473	intanm/fin-sa-post-trained-indobert-large-finnews-p2-002
M424	intanm/mlm_v1_20230327_fin_sa_60	M474	jordankrishnayah/robertaSentimentBiasTestwithMBIC
M425	raygx/distilGPT-NepSA	M475	adnanakbr/emotion-english-distilroberta-base-fine_tuned_for_amazon_english_reviews_on_200K_review
M426	gsl22/sentiment-analysis-v1	M476	BeChi87/train_model_noStopWord
M427	intanm/100-finSA-finLM-v1-001	M477	adnanakbr/emotion-english-distilroberta-base-fine_tuned_for_amazon_english_reviews_on_200K_review_v3
M428	intanm/baseline_fin_sa_100	M478	intanm/fin-sa-post-trained-indobert-base-finreports-p2-001
M429	venetis/deberta-v3-base-finetuned-3d-sentiment	M479	venetis/distilroberta-base-finetuned-3d-sentiment
M430	manvik28/FinBERT_Tuned	M480	gabrielkytz/novo
M431	NewtonKimathi/Covid_Vaccine_Sentiment_Analysis_Bert_based_Model	M481	sshs/phobert-base-vietnamese-sentiment
M432	raygx/GNePT-NepSA	M482	BeChi87/train_model_Lastest_new
M433	santis2/phrasebank-sentiment-analysis	M483	jordankrishnayah/fullManifesto-ROBERTAsentiment2e
M434	venetis/electra-base-discriminator-finetuned-3d-sentiment	M484	BeChi87/train_model_Lastest_21
M435	AlonCohen/RuSentNE-iter-2	M485	BeChi87/train_model_Lastest_16_8_lan2
M436	intanm/baseline_fin_sa_70	M486	BeChi87/train_model_Lastest_16_8_lan3
M437	jcy204/cold_model	M487	maegancp/finetuning-sentiment-model
M438	Beanz1935/finetuning-sentiment-model-3000-samples	M488	mrfakename/distilroberta-financial-news-tweets-sentiment-analysis
M439	yanezh/twiiter_try15_fold2	M489	Jonathan0528/bert-base-uncased-financial-news-sentiment
M440	intanm/baseline_fin_sa_30	M490	rhythm00/HF_tutor
M441	Robher51/Model3	M491	salohnana2018/ABSA-Pair-domainAdapt-CAMEL-MSA-HARD-ImbalancedHandling-absoluteWeights-gamma1-dropout01
M442	yanezh/twiiter_try15_fold4	M492	salohnana2018/ABSA-Pair-domainAdapt-CAMEL-MSA-HARD-ImbalancedHandling-inv-prop-weights-gamma1
M443	tiwanyan/distil_bert_fine_tune	M493	nuriafari/my_model
M444	intanm/fin-sa-post-trained-indobert-base-finnews-5	M494	salohnana2018/ABSA-SentencePair-DAPT-HARD-SemEval-bert-base-Camel-MSA-run3
M445	yanezh/twiiter_try15_fold3	M495	salohnana2018/ABSA-Pair-domainAdapt-Qarib-HARD-ImbalancedHandling-absoluteWeights-gamma1-dropout01
M446	cruiser/distilbert_final_config_dropout	M496	salohnana2018/ABSA-Pair-domainAdapt-Camel-MSA-HARD-ImbalancedHandling-absoluteWeights-gamma1-run2
M447	Q317/EmoraBert4	M497	salohnana2018/ABSA-Pair-domainAdapt-Camel-MSA-HARD-ImbalancedHandling-absoluteWeights-gamma1-fulldata
M448	raygx/xlmRoBERTa-NepSA	M498	salohnana2018/ABSA-Pair-domainAdapt-Camel-MSA-HARD-ImbalancedHandling-absoluteWeights-gamma1-run1
M449	yanezh/twiiter_try10_fold0	M499	salohnana2018/ABSA-Pair-domainAdapt-CAMEL-MSA-HARD-ImbalancedHandling-norm_inv_const5-gamma3
M450	yanezh/twiiter_try11_fold3	M500	Anwaarma/autotrain-testsen-65460136124

Table 11: Split-5 of M_{500} Models

A.2 In-Domain/Out-Domain Evaluation of M_{500}

Since most of the model authors have not reported their training datasets, it becomes difficult to determine whether the model evaluation on a test dataset is in-domain or out-domain. This is the reason we avoided conducting an in-domain/out-domain evaluation. However, in this section, we present the results (refer Table 12) based on the available information for the models in their model cards. The results may not be conclusive since the training data for most models in M_{500} is unknown. Furthermore, as depicted in Figure 8 and similar figures that follow, it is evident that performance is not solely dependent on whether the evaluation is In-Domain or Out-Domain. For instance, Figure 8 illustrates that some models in Out-Domain evaluation perform better than models in In-Domain evaluation.

A.3 Granular analysis of model card sections

A more granular analysis extending §4.2 with respect to sections of model cards is presented in Table 13.

B Annotator Information

We recruited annotators for two tasks in our study. Task 1 involved the Manual Inspection study outlined in §3.2, while Task 2 focused on curating the Reddit dataset as described in §4.3. We invited undergraduate students to volunteer for these tasks, ensuring their anonymity and providing no compensation. Detailed annotation instructions can be found in §E for Task 2. Annotators were not given strict time limits for annotation. They were proficient in English and familiar with the tasks. For Task 2, in the event of a tie between annotator labels, we randomly selected a label based on their order of labeling as the final decision.

C Hyperparameters

We did hyperparameter tuning on RoBERTa model for all datasets in $\mathcal{D}$, the optimal parameters are detailed in Table 6. The optimization process employs the AdamW optimizer (Loshchilov and Hutter 2019). To ensure reproducibility, we conduct all our experiments using the open-source HF Transformers library[40] (Wolf et al. 2020). Additionally, all experiments are executed using $4 \times$ Tesla V100 GPUs.

D Performance Plots

Refer to Figure 8 and other subsequent plots.

E Annotation Instructions

Refer to Figure 13 for annotation instructions.

[40] https://huggingface.co/docs/transformers/

		$\mathcal{D}_t$						$\mathcal{D}_a$						$\mathcal{D}_f$					
		In-Domain			Out-Domain			In-Domain			Out-Domain			In-Domain			Out-Domain		
		Easy	**Ambi**	**Hard**	**Easy**	**Ambi**	**Hard**	**Easy**	**Ambi**	**Hard**	**Easy**	**Ambi**	**Hard**	**Easy**	**Ambi**	**Hard**	**Easy**	**Ambi**	**Hard**
Evaluate	r_d	0.058	0.235	0.271	0.214*	0.186*	0.057	-0.714	-0.714	-0.642	0.217*	0.188*	0.162*	0.113	-	-	0.107*	-	-
	r_l	-0.035	0.051	0.250	0.159*	0.121*	0.077	-0.883*	-0.883*	-0.270	0.140*	0.154*	0.173*	0.230	-	-	0.153*	-	-
	r_{da}	0.105	0.146	0.153	-0.098*	-0.089	-0.063	0.178	0.178	0.393	-0.084	-0.081	-0.135*	-0.576*	-	-	-0.086	-	-
Finetune	r_d	0.018	0.34	0.075	-0.048	-0.115*	0.093	-0.535	-0.535	0.107	-0.087	-0.068	0.151*	0.268	-	-	-0.098*	-	-
	r_l	-0.097	0.074	0.083	-0.162*	-0.187*	0.175*	-0.558	-0.883*	0.324	-0.161*	-0.168*	0.247*	0.263	-	-	-0.128*	-	-
	r_{da}	0.099	0.147	0.281	0.047	0.031	-0.122*	0.642	0.321	0.393	-0.011	-0.015	-0.052	-0.433*	-	-	0.009	-	-

Table 12: Correlation between r and the performance (F1) of models in M_{500} across different dataset domains and categories. Values highlighted with * represent statistical significant values with p-value < 0.05. For $\mathcal{D}_f$, we did not perform any evaluation strategy for ambiguous and hard instances, since the #instances categorized as ambiguous and hard is very low, which is insufficient for training and evaluation, therefore values are not calculated and displayed.

		$\mathcal{D}_a$			$\mathcal{D}_a$			$\mathcal{D}_f$	r		
		Easy	**Ambi**	**Hard**	**Easy**	**Ambi**	**Hard**	**Easy**	r_d	r_l	r_{da}
Model Details	**FIA**	0.173*	0.159*	0.064	0.132*	0.139*	0.135*	0.102*	0.603*	0.521*	-0.039
	PIA	0.076	0.054	-0.051	0.038	0.046	0.047	-0.006	0.504*	0.408*	-0.159*
	NIA	-0.189*	-0.173*	-0.088	-0.139*	-0.160*	-0.180*	-0.139*	-0.667*	-0.521*	0.079
Training Details	**FIA**	0.073	0.063	0.020	0.097*	0.067	0.017	0.074	0.384*	0.272*	0.112*
	PIA	0.052	0.038	-0.055	0.047	0.027	0.030	0.074	0.343*	0.335*	-0.152*
	NIA	-0.079	-0.078	-0.015	-0.110*	-0.085	-0.038	-0.105*	-0.448*	-0.327*	-0.019
Evaluation Details	**FIA**	0.119*	0.133*	0.048	0.093*	0.101*	0.125*	0.090	0.199*	0.147*	-0.081
	PIA	0.099*	0.078	-0.069	0.073	0.051	0.032	0.083	0.289*	0.307*	-0.102*
	NIA	-0.165*	-0.160*	-0.011	-0.134*	-0.140*	-0.116*	-0.152*	-0.305*	-0.226*	0.067
Uses	**FIA**	0.137*	0.112*	-0.082	0.109*	0.088	0.087	0.049	0.501*	0.444*	-0.171*
	PIA	0.118*	0.090	-0.079	0.098*	0.082	0.039	0.128*	0.278*	0.280*	-0.118*
	NIA	-0.176*	-0.149*	0.068	-0.144*	-0.121*	-0.119*	-0.094*	-0.458*	-0.444*	0.195*
Other Details	**FIA**	0.146*	0.128*	-0.015	0.131*	0.114*	0.081	0.070	0.411*	0.334*	-0.173*
	PIA	0.079	0.060	-0.043	0.062	0.047	0.060	0.110*	0.327*	0.344*	-0.088*
	NIA	-0.028	-0.019	0.088	-0.024	-0.009	0.011	-0.033	-0.068	-0.095*	0.012

Table 13: Correlation values computed between the percentage of FIA/PIA/NIA (across all sections of the model card) and model performance (F1). Additionally, it includes correlation values calculated between the percentage of FIA/PIA/NIA and r (model popularity). Values highlighted with * represent statistical significant values with p-value < 0.05. For $\mathcal{D}_f$, we did not perform any evaluation strategy for ambiguous and hard instances, since the #instances categorized as ambiguous and hard is very low, which is insufficient for training and evaluation, therefore values are not calculated and displayed.

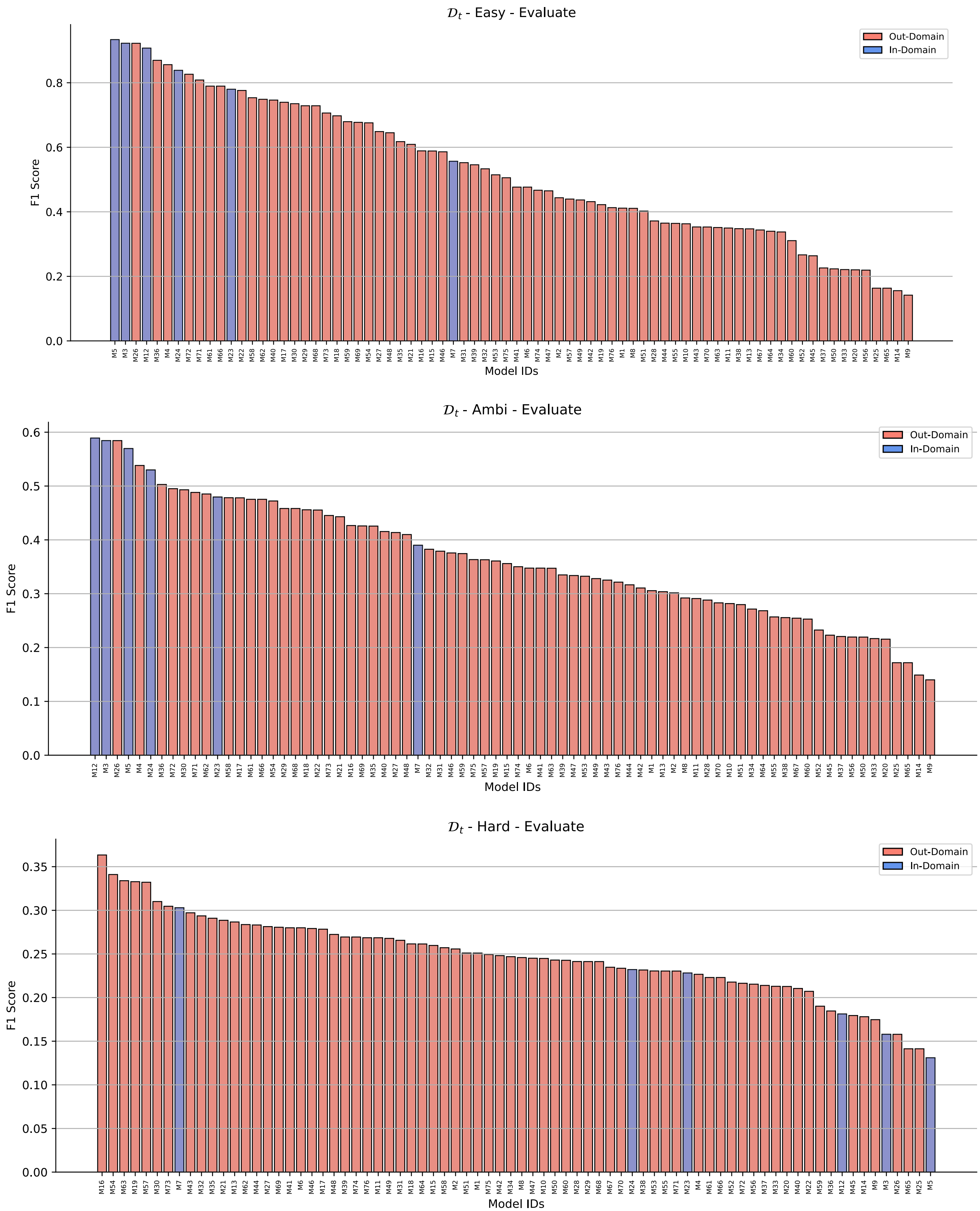

Figure 8: TweetEval – Evaluate

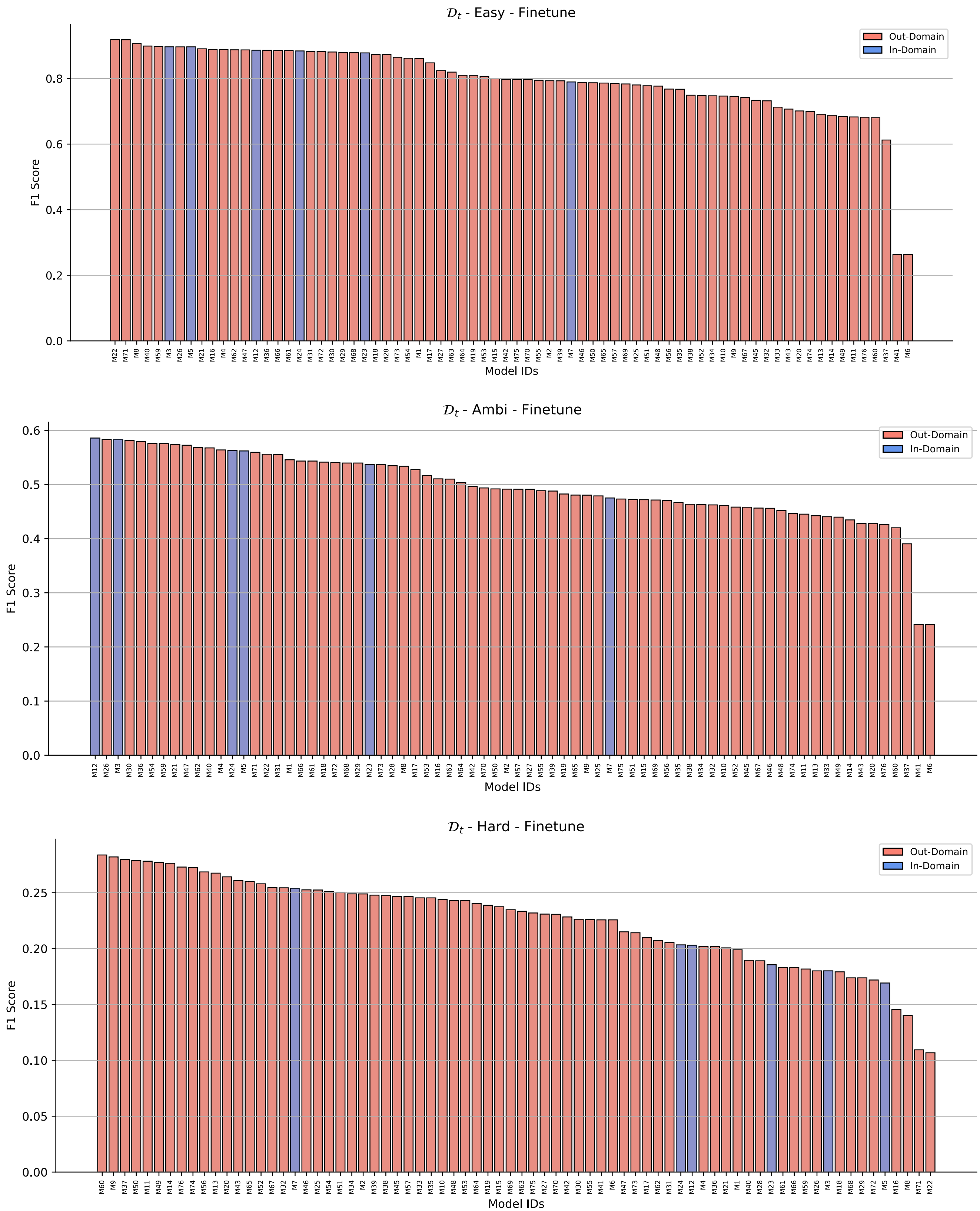

Figure 9: TweetEval – Finetune

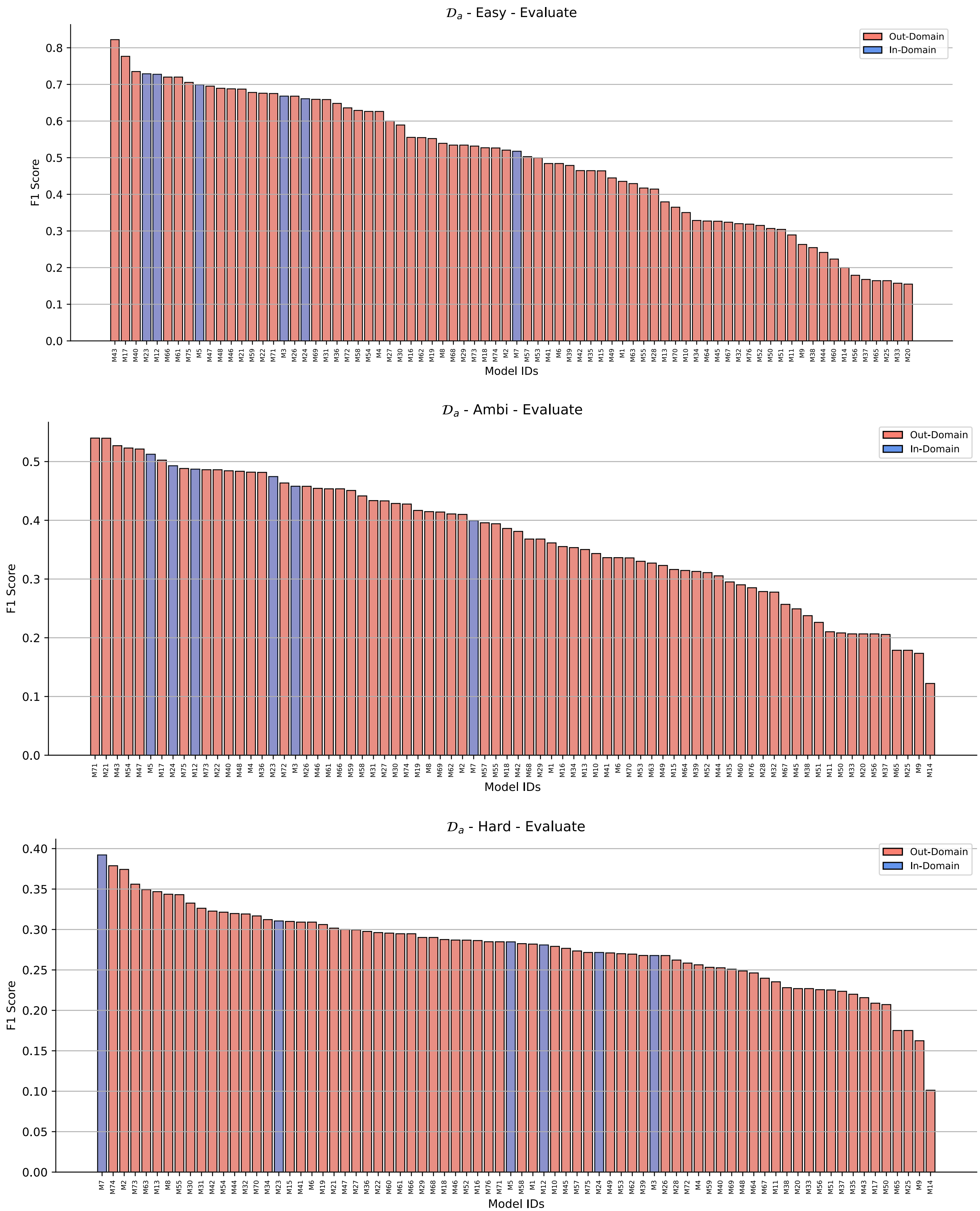

Figure 10: Amazon Multi Reviews – Evaluate

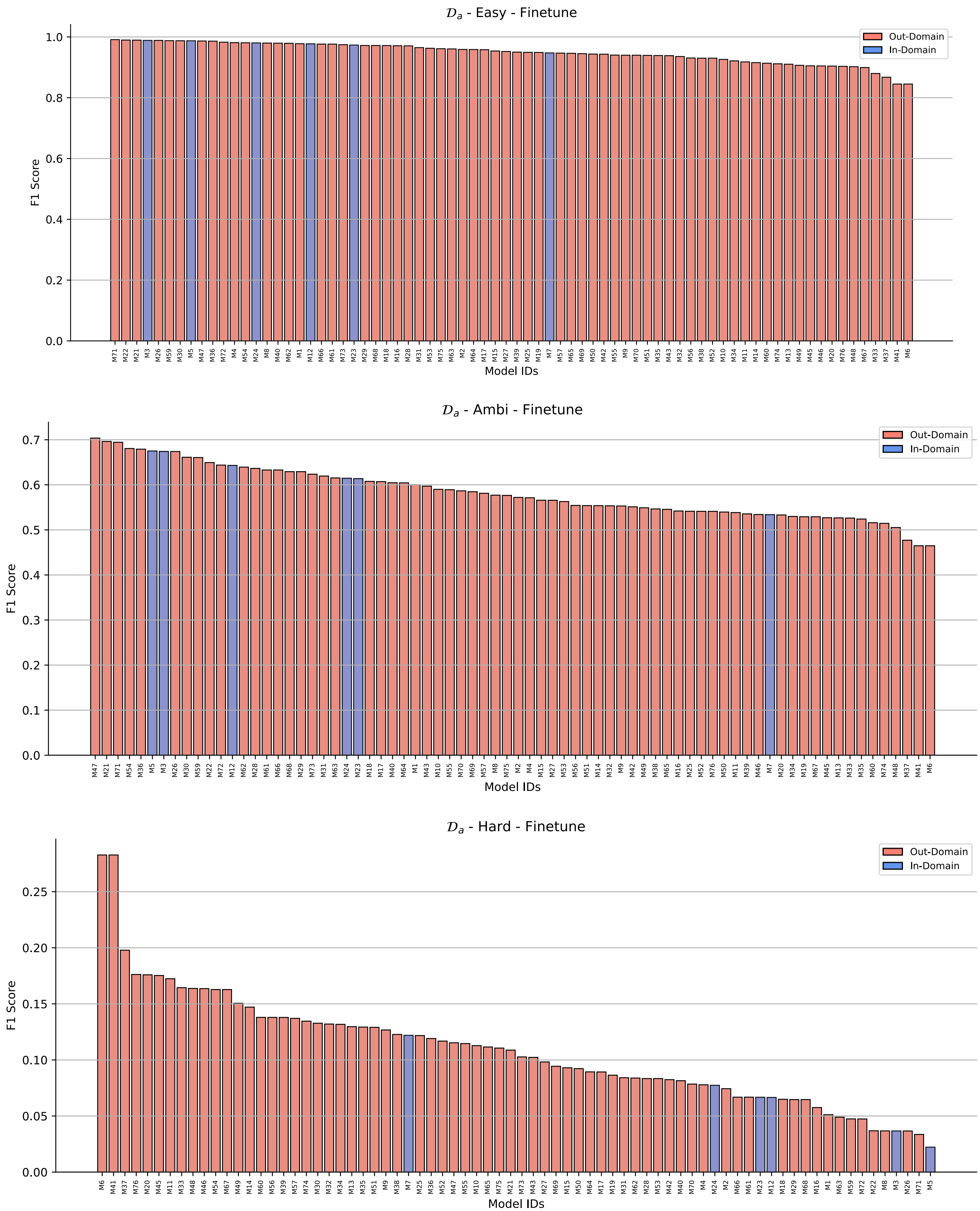

Figure 11: Amazon Multi Reviews – Evaluate

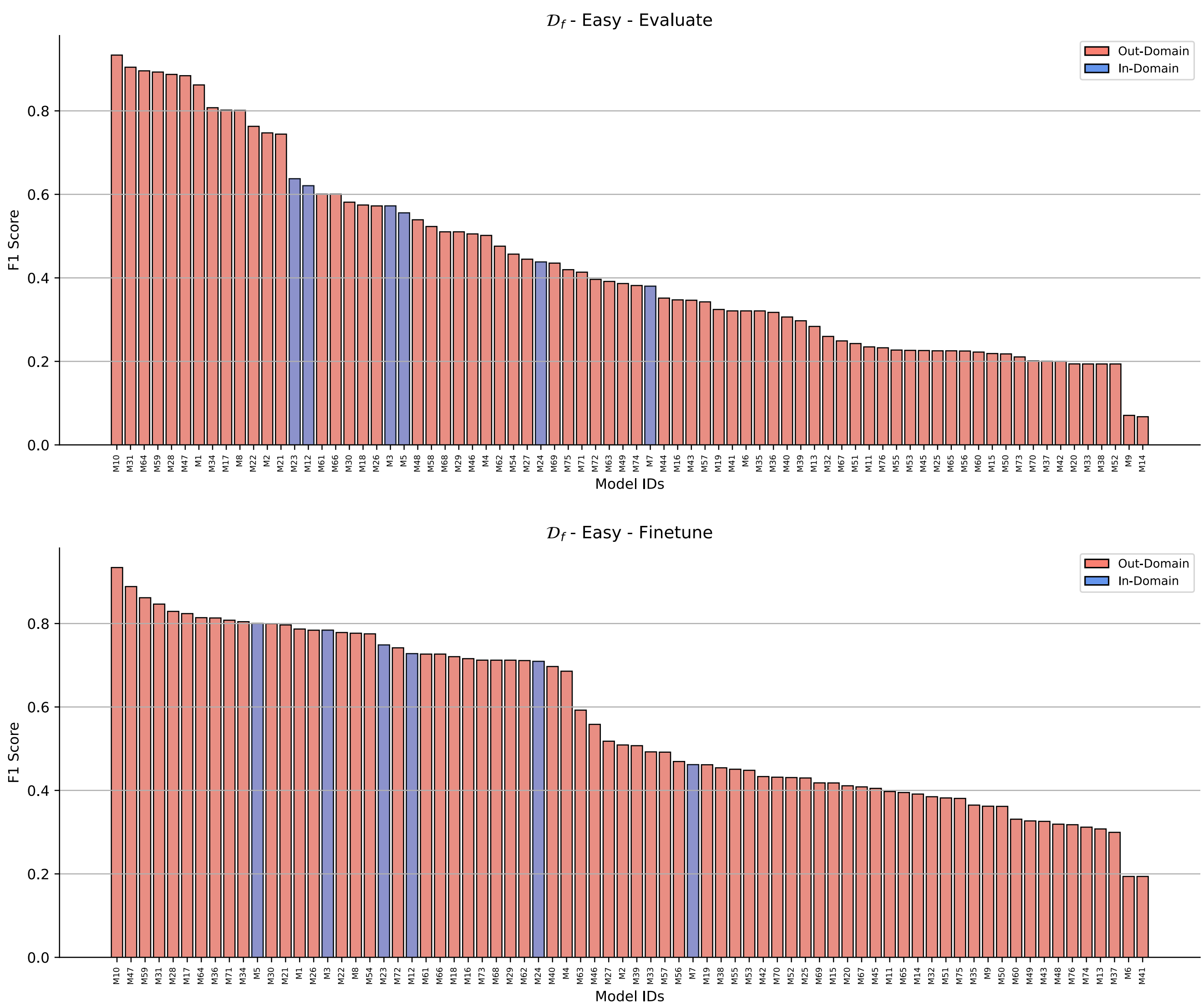

Figure 12: Financial Phrasebank – Evaluate – Finetune

Task: Sentiment Analysis Annotation

Objective:

Your task is to analyze and annotate the sentiment expressed in text samples. The sentiment can be positive, negative, or neutral. Label the text as 2 for Positive sentiment, 0 for negative sentiment and 1 for Neutral sentiment. If you are unsure of any of the sentiment, label it with the most appropriate label and mark it as 1 for Uncertain, you just have to flip the label from 0 to 1, check the google sheet, you will understand. Please follow the guidelines below for accurate and consistent annotation.

You are expected to annotate ~100 samples in a day.

Instructions:

Familiarize Yourself with Sentiment Classes:

2: **[Positive],** The text expresses a positive sentiment. The speaker/author is happy, satisfied, or optimistic.
0: [Negative], The text expresses a negative sentiment. The speaker/author is unhappy, dissatisfied, or critical.
1: [Neutral], The text does not convey a particular sentiment. It is objective, factual, or does not express emotion.

Consider the Context:

Take into account the context of the entire text when determining sentiment.
Consider the tone, language, and overall message to accurately capture the sentiment.

Single Sentiment Label:

Assign a single sentiment label to each text sample.
If a text contains both positive and negative sentiments, choose the predominant sentiment.

Be Mindful of Ambiguity:

If the sentiment is unclear or ambiguous, mark it as neutral.
Do not guess the sentiment; only annotate what is explicitly expressed in the text.

Handle Sarcasm/Irony:

If the sentiment is conveyed through sarcasm or irony, annotate based on the intended sentiment, not the literal meaning.

Figure 13: Annotation Instructions (Page 1)

Handle Mixed Sentiments:

If a text expresses both positive and negative sentiments, choose the overall sentiment that is more prevalent.

Skip Irrelevant Text:

If the text does not express a sentiment or is irrelevant to sentiment analysis, mark it as neutral.

Report Technical Issues:

Report any technical issues or difficulties encountered during annotation.
Do not hesitate to ask for clarification if the instructions are unclear.

Quality over Quantity:

Prioritize accuracy in your annotations over speed.
Take the time to carefully read and understand each text before assigning a sentiment label.

Example Annotations:

Text: "The new phone is amazing, but the battery life is disappointing."
Annotation: Negative

Text: "The customer service experience was terrible; I waited on hold for over an hour and received no resolution to my issue."
Annotation: Negative

Text: "After watching the movie, I was disappointed by the weak plot and unconvincing performances from the actors."
Annotation: Negative

Text: "The report presents the facts and figures without expressing any opinion."
Annotation: Neutral

Text: "The weather forecast predicts clear skies and moderate temperatures for the weekend."
Annotation: Neutral

Text: "This manual provides step-by-step instructions for assembling the furniture."
Annotation: Neutral

Positive Sentiment:
Text: "I just received a promotion at work! Feeling ecstatic and grateful for this opportunity."

Figure 14: Annotation Instructions (Page 2)

Annotation: Positive

Text: "The team worked collaboratively to achieve the project goals, and the results exceeded expectations."
Annotation: Positive

Text: "Had an amazing time at the concert last night. The performance was outstanding, and the crowd was lively!"
Annotation: Positive

Figure 15: Annotation Instructions (Page 3)

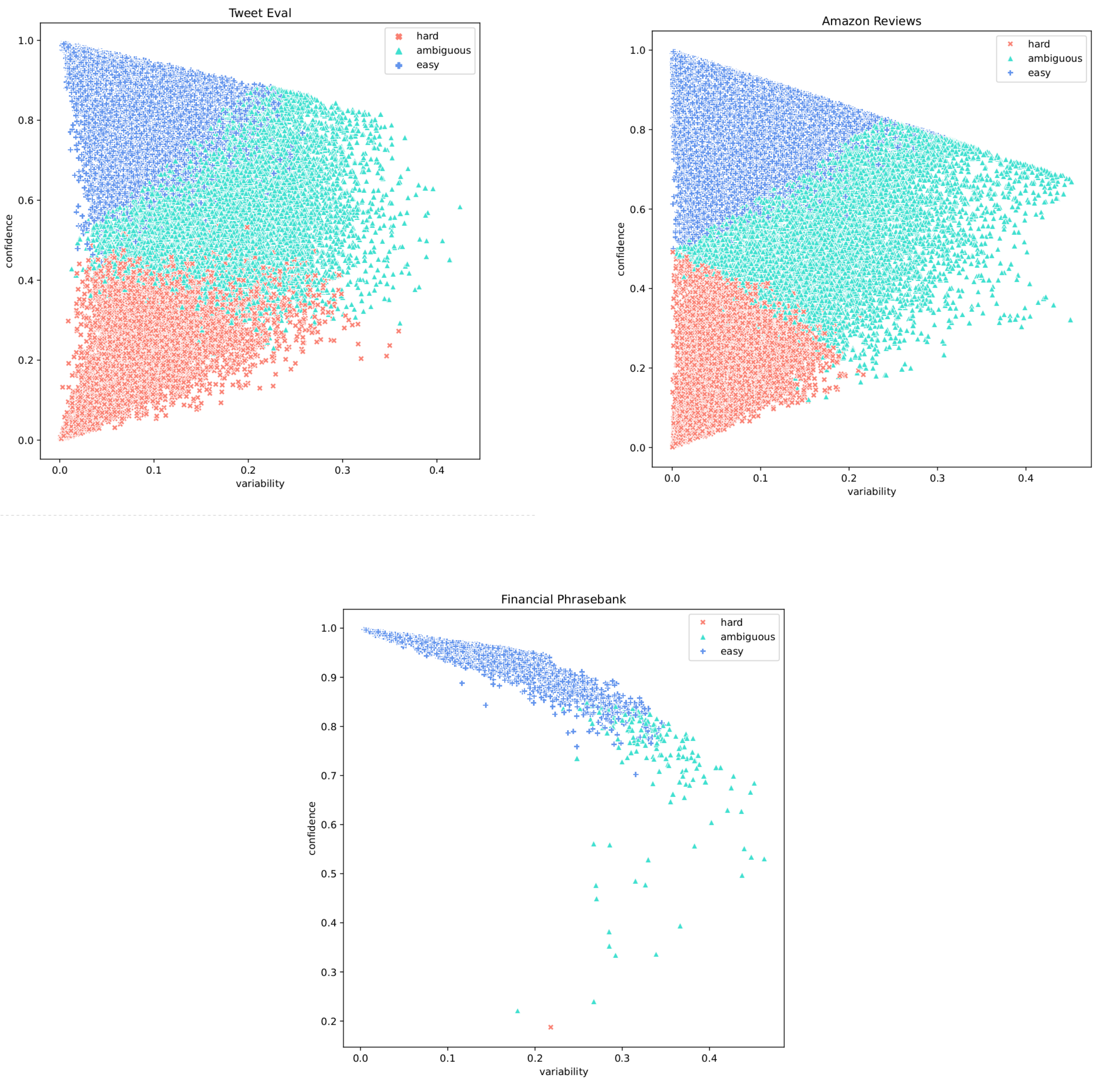

Figure 16: Datamaps across $\mathcal{D}$