Enhancing Sentiment Analysis Through Transfer Learning: Adapting BERT, RoBERTa, and DistilBERT Architectures Across Varied Domains Submissions

Anonymous TACL submission

Abstract

This research explores the advancement of binary sentiment analysis through the application of transfer learning on leading language models such as BERT, RoBERTa, and DistilBERT. Focused on three key domains (e-commerce, entertainment and financial markets) this study leverages domain-specific datasets to fine-tune these models, aiming to improve their sentiment detection capabilities. The investigation reveals significant increase in weighted F1 scores over conventional baseline models, and with some DistilBERT and BERT maintaining consistent best performances across different datasets, showcasing the adaptability and efficiency of each model across different sectors. By emphasizing the benefits of model adaptation within natural language processing, this study not only highlights the effectiveness of transfer learning in sentiment analysis tasks but also offers insights into the strategic selection and optimization of language models for better performance across varied domains. This approach underscores the potential for broader applications of these findings in enhancing sentiment analysis.

1 Introduction

The rise of technology-driven platforms and websites has led to a great increase in user-generated data, as users have increased freedom to express their opinions on various avenues (Do et al., 2019). These opinions range from reviews on the food and beverages sector, e-commerce items bought, and other services, to social media such as Facebook, and Reddit. User-generated data stems from the behaviour and decision-making process of customers or opinionated individuals, and they can be extremely useful in helping us understand the underlying motivation for making such remarks. However, manually collecting user-generated data is a monumental task due to the sheer enormous amount of data there is out there, and is unfeasible because of time or financial constraints. Sentiment analysis is a method of providing insights into user content by analyzing people's opinions and emotions towards an event or situation. Zhang et al. (2018) discussed sentiment analysis from papers from three levels: aspect level, sentence level, and document level. These order the granularity of text analysis, and provide a "target". User-generated content usually consists of two parts: the "target", and the "sentiment". For example, a review such as: "The fish and chips (target) are terrible (sentiment)" contains both parts, and it is the task of sentiment analysis to identify these parts within a given text.

Lagrari and Elkettani (2021) places traditional sentiment analysis into two types: lexicon-based and machine learning-based. Lexicon-based methods describe the method of labeling the document based on a sentiment lexicon, which is a list of predetermined sentiment polarity of phrases and words, by summing up the polarity score. Machine Learning approaches describe building a model by learning from labeled datasets, and using the model to classify sentiment of a given input. Popular methods include Naive Bayes or Support Vector Machines (SVMs). These approaches, however, require large amounts of data to reduce overfitting, and to improve the generalization of the model. Such data can be difficult or expensive to obtain in certain domains. As such, transfer learning in NLP (Pan and Yang, 2009) has gained attention in leveraging existing data within these domains to improve the performance of similar tasks.

In our study, we explore the technique of transfer learning. We tested the performance of three

different models: BERT, RoBERTa, and DistilBERT on four datasets. The main goal is to investigate if these models perform well when specific datasets are used for fine tuning, whether order of fine tuning affects the performance, and if models remain robust across different levels of finetune.

2 Related Works

Enhancing sentiment analysis through transfer learning and domain adaptation has become a focal point in leveraging the capabilities of Large Language Models (LLMs) across varied domains. The integration of novel methodologies and the critical evaluation of these approaches offer insights into both the potential and the limitations of LLMs for nuanced sentiment analysis.

Du et al. (2020) introduce a pioneering method to mitigate domain discrepancy in cross-domain sentiment analysis, enhancing BERT's applicability across distinct domains. Their two-stage strategy incorporates a domain-distinguish pre-training task, equipping BERT with domain recognition capabilities, followed by adversarial training to facilitate knowledge transfer to less resource-rich domains. This methodology not only showcases superior performance over existing approaches but also opens avenues for further research on alternative pre-training tasks, unsupervised learning techniques, and model explainability in cross-domain settings.

Tan et al. (2022) introduce F-OTCE and JC-OTCE metrics that are used for cross-domain cross-task transfer learning which significantly advances the field by eliminating the need for auxiliary tasks in transferability evaluation, thus streamlining the process. The utilization of optimal transport for assessing Negative Conditional Entropy between source and target labels enhances both efficiency and accuracy, setting a new standard for evaluating transferability. However, the research also highlights areas needing attention: the computational intensity required by JC-OTCE and its reliance on specific solutions could limit broader application. The paper effectively balances these innovative strengths with candid acknowledgments of potential weaknesses, laying a solid foundation for future explorations to optimize these metrics further and expand their applicability across more diverse datasets and learning scenarios. This nuanced approach underscores the complexity of transfer learning challenges and the ongoing need for refined solutions.

(Shingu et al., 2021) embarks on the critical task of deciding on the variables that predict the success of domain adaptation in the realm of text similarity. Utilizing descriptive domain information and cross-domain similarity metrics, the research posits a framework to foresee the effectiveness of applying models trained in one domain (source) to another (target). Despite the innovative approach, the study's exploration into the robustness of these predictive features across varied real-world scenarios remains limited, potentially overlooking the complexities and biases inherent in diverse datasets. Moreover, the absence of a thorough comparison with existing domain adaptation methodologies restricts the ability to contextualize the performance and innovation of the proposed model within the broader landscape of the field. The paper's recommendations for future research, while well-intentioned, lack specificity, offering little in terms of a concrete strategy or explicit directions for further investigation. Furthermore, the reliance on a select group of datasets for validation raises questions about the model's generalizability and applicability across different domains and tasks. This critique not only highlights the significant contributions of the paper to understanding domain adaptation in text similarity tasks but also underscores the necessity for a more comprehensive and comparative analysis, a detailed exploration of feature robustness, and an expansion of validation across a wider array of domains to truly advance the field.

The pursuit of advanced sentiment analysis methods must also consider the ethical implications of deploying LLMs, including the potential for bias amplification and the manipulation of public opinion. As such, ongoing development in LLM architectures and the quest for more comprehensive evaluation methodologies are essential.

Collectively, these studies give valuable perspectives on the challenges and opportunities in applying LLMs and transfer learning techniques for sentiment analysis across a spectrum of domains. By addressing the highlighted limitations and exploring the suggested future research directions, the field can progress towards realizing the full potential of LLMs in revolutionizing sentiment analysis.

While many papers address the concept of both transfer learning tasks and base robustness across

models individually, there is a lack of discussion about the combination of the two. This can range from the exploration of how models that have been trained on a diverse set of tasks through transfer learning can exhibit different levels of robustness when faced with adversarial attacks, to the effectiveness of these models in retaining their performance across various domains or under distribution shifts. This gap in the literature highlights the need for comprehensive studies that not only investigate the performance of transfer-learned models in their target tasks but also scrutinize their robustness.

3 Methods

3.1 Research question

In this paper, we address a niche field that inspire new models that are able to specialise in adaptability and robustness across a variety of domains and implementations. Specifically, we are looking at how adaptable and robust are finely-tuned machine learning models on a various semantic domains.

Most literature defines robustness as model robustness, in which fine-tuning is measured against adversarial tasks (shafahi et al.2020). Following this, such papers introduce custom large datasets with adversarial permutations in a methodological way, (Moosavi-Dezfooli et al, 2016). The robustness is usually evaluated from attack effectiveness (i.e., attack success rate). While we do touch on some of these concepts, robustness in this context is defined similarly to versatility and transferability. Tan et al, 2024 act on this concept in detail uncovering two new metrics to evaluate it. While a strict definition of robustness cannot be given, we can derive the definition from the question at hand: i.e how adaptable are models to a variety of domains via the usage of transfer learning? Intuitively, we define our datasets and metrics in accordance with this, emphasising the adaptability of machine learning models when applied across various semantics fields. By doing this, it allows us to review the nuanced capabilities of the machine learning models whilst measuring their effectiveness in dynamic environments and transferability in real-world scenarios.

3.2 Data Collection

We used four distinct datasets, each representative of varied domains, including a specially curated adversarial dataset designed to challenge the models with complex linguistic expressions such as sarcasm, indirectness, and subtlety. Opting for binary sentiment classification tasks, focusing solely on distinguishing between positive and negative sentiments, not only simplifies the evaluation framework but also mitigates the challenge of varying sentiment categories across datasets. This approach ensures a more consistent and scalable application of the BERT, GPT, and RoBERTa language models across diverse contexts, including entertainment, finance, e-commerce, and intentionally ambiguous scenarios presented by the adversarial dataset. The binary classification scheme, alongside the inclusion of adversarial examples, enables straightforward comparisons, transferability of insights, and particularly effective cross-domain sentiment analysis by challenging the models to decipher nuanced language beyond domain-specific keywords, thus aiming for a more nuanced understanding of sentiment. The size of the datasets are shown in Table 1t.

Evaluation Dataset	No. of Sentences
Amazon	36000
IMDB	25000
Financial Phrasebank	1965
SST-2	9613
adversarial	50

Table 1: Summary of datasets used.

3.2.1 IMDB Movie Reviews

The IMDB movie reviews dataset, hosted on Hugging Face, comprises 50,000 reviews from the IMDB database. This dataset is evenly split into 25,000 reviews for training and 25,000 for testing, with an equal distribution of positive and negative reviews. Its binary nature, distinguishing between positive and negative sentiments, makes it an ideal candidate for fine-tuning language models to understand nuances in consumer sentiment in the entertainment industry.

3.2.2 Financial PhraseBank

The Financial PhraseBank dataset contains sentences from financial news articles, classified into positive, neutral, or negative sentiments. The dataset was adapted to a binary classification task by excluding neutral instances, focusing on the polarity of sentiments expressed in financial texts.

We utilized the "sentences_all_agree" subset, ensuring only sentences with unanimous agreement among annotators were included. This selection criterion is crucial for reliably evaluating the model's performance in understanding complex, domain-specific jargon and sentiment in the financial sector.

3.2.3 Amazon

The Amazon product reviews dataset, available on Hugging Face, consists of reviews collected from the Amazon website. To reduce training time, we opted to use 1% of the full dataset. This subset includes reviews with binary sentiment labels, dividing them into positive and negative categories. It is structured with an even distribution of positive and negative reviews, making it a balanced dataset for training and testing natural language processing models.

3.2.4 Stanford Sentiment Treebank (SST)

SST-2 is a dataset commonly used for sentiment classification which contains two classes, and contains generalized sentences.

3.2.5 Adversarial Dataset

The adversarial dataset was manually curated to challenge sentiment analysis models by featuring a range of complex linguistic expressions, such as sarcasm, indirectness, and subtlety, marked with binary sentiment labels. The entries encompass a broad range of contexts, including reviews of hotels, books, movies, personal experiences, and environmental concerns, intentionally employing language that often contradicts the sentiment conveyed, such as positive phrases with negative intentions. This collection tests and enhances the models' capabilities to decipher nuanced language and accurately classify sentiments that go beyond domain-specific keywords. The full dataset and example of entries can be found in the appendix, A.

3.3 Models

We opt to perform our experiments on three popular models that have been shown to achieve state-of-the-art performance on a range of NLP tasks, namely, BERT (Devlin et al., 2018) RoBERTa, (Liu et al., 2019) and DistilBERT (Sanh et al., 2019). The models not only have been pre-trained on large and diverse text corpora, but they also represent different architectures. For example, BERT uses a bidirectional transformer, RoBERTa is an optimized version of BERT with changes in pre-training procedures, and DistilBERT is a distilled version of BERT that is smaller and faster while retaining most of its performance.

3.4 Transfer-learning

The form of transfer learning we apply is typically known as *fine-tuning* (Houlsby et al., 2019). In this case, we can formally define it as continued training with pre-defined weights to obtain adjusted weights to obtain which is optimized for the new dataset. Mathematically, we utilise:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(x_i; \theta), y_i) \tag{1}$$

Where

- $L(\theta)$ represents the loss function computed over the parameters θ.
- The fraction $\frac{1}{N}$ denotes the mean over the N samples.
- The summation $\sum_{i=1}^{N}$ iterates over each sample.
- $\mathcal{L}$ denotes the loss for each prediction compared to the true label y_i. In our case, we utilise binary cross-entropy loss due to our binary labels
- $f(x_i; \theta)$ represents the model's prediction for the input x_i with parameters θ.

Furthermore, when finetuning we do not freeze any layers, rather we adjust the weights across all layers of the pre-trained model to better fit the data from the specific domain. This can lead to improved performance on tasks related to those domains, as the model adjusts from the general language understanding learned during pre-training to the specifics of the new data.

The chosen hyperparameters are shown table 3.4. These are optimized for fine-tuning a general pre-trained model on a new task. A small learning rate of 5×10^{-5} is used to make small, precise updates without overriding pre-learned features, and the AdamW optimizer for its balance between fast convergence and effective regularization. A batch size of 8, for both training and evaluation, ensures a balance between computational efficiency and model stability, while limiting the training to

3 epochs minimizes the risk of overfitting and is common for fine-tuning tasks.

Table 2: Hyperparameters

Learning rate	5×10^{-5}
Optimizer	AdamW
Batch size (per device, train)	8
Batch size (per device, eval)	8
Number of epochs	3

3.5 Metrics

Weiss and Khoshgoftaar (2017) provides a well-done analysis of different metrics to measure transfer learning techniques. The literature often refers to the commonly used metric of AUC which is great for our binary classification models, but still poses some problems. For instance, AUC can be overly optimistic in cases of severe class imbalance. The ROC curve, and consequently the AUC, might suggest good performance even when the model has a significant bias towards the majority class. Instead, the literature also opts for the ROC Curves in combination with AUC scores. Taking this into consideration we utilise these metrics to provide a correct robustness calculation:

- **Macro/Micro-Averaged F1 Score:** In sentiment analysis, you may have imbalanced classes, where one sentiment class dominates the dataset. In such cases, macro and micro-averaged F1 scores provide a better understanding of model performance across all classes. Micro-averaged F1 score calculates metrics globally by counting the total true positives, false negatives, and false positives, while macro-averaged F1 score calculates metrics for each label and then takes the average, giving equal weight to each class.

- **Weighted F1 Score:** Similar to macro and micro-averaged F1 score, but it calculates the F1 score for each class and then takes the average, with each class weighted by its proportion in the dataset. This is useful when you have imbalanced datasets.

- **Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC):** While ROC curves and AUC are typically used in binary classification tasks, they can also be applied in sentiment analysis, especially when predicting sentiment as positive or negative. ROC curves visualize the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity), and AUC quantifies the model's ability to discriminate between positive and negative sentiments

We decided to opt for weighted F1-scores in this case, due to 2 reasons: class imbalance, and simplicity. The size of the datasets varies greatly from just over 2k in the Financial Phrasebank and 36k in Amazon Polarity, with class imbalance within the datasets; This calls for a metric that normalizes the metric, and provides a representative insight into the performance. The analysis aims to focus on the comparison of robustness of models across different combinations of fine tune, hence the metric should be one that is simple and easy to interpret, like that of the weighted F1-scores.

4 Experiments

Our hypothesis is that models will do worse when they are fine tuned and evaluated over different domains (e.g. fine tuned on Amazon Polarity and evaluated on Financial Phrasebank). We seek to understand this by iterating all 3 models through permutations of 4 different datasets, and employing the Leave-One-Out (LOO) approach. As we are also experimenting on different orders of fine tuning, there are in total 24 methods tested (4 combinations for 4 datasets, and multiplied by 6 different arrangements in each combination). The full pipeline can be found in the GitHub repository at A.

4.1 Base models

The infrastructure that we utilise in our base models is carried forth in the fine-tuning of the models. The infrastructure we employ involves firstly calculating certain metrics utilising just the default models. We benchmark these models with the parameters highlighted in table 3.4 to acquire the weighted F1-scores.

4.2 Leave-One-Out Cross Validation

We utilise the technique of Leave-One-Out (LOO) cross validation to reduce bias across all the data, and to assess the robustness of the individual models. While we are not strictly using cross validation within the training data of each dataset, we are using each dataset as a form of cross validation. This ties in with our goal of examining the robustness of each model.

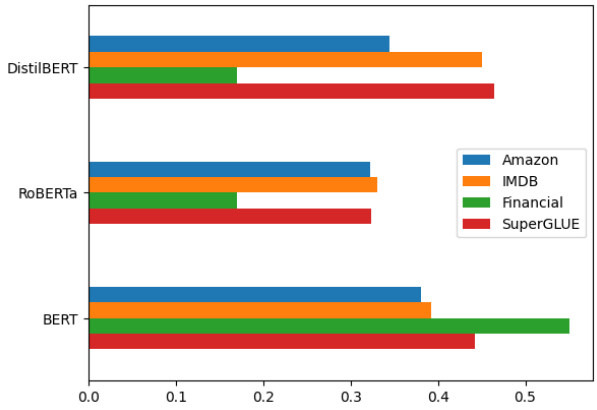

Figure 1: Comparison of weighted F1-scores of baseline results across DistilBERT, RoBERTa, and BERT.

5 Results

The results section will outline 4 different points: (1) Baseline results, (2) Effectiveness of transfer learning, (3) Robustness of models across fine tuning, and (4) Impact of fine tuning order.

5.1 Baseline Results

The performance of the baseline system, on each of the 4 datasets, is shown in Fig. 1. The main observation is that overall, BERT seems to do the best, achieving on average an weighted F1-score of about 0.45, while RoBERTa seems to do the worst, having an average of around 0.25. As we used the same dataset across all three models, it suggests that an optimised model such as RoBERTa performs worse when we feed it insufficient data, as it is data hungry and require significantly more data to perform as expected. Another interesting insight is that both DistilBERT and RoBERTa seem to do significantly worse than BERT on the Financial Phrasebank dataset, which seems to suggest that optimised or more efficient models could impede their performance when evaluated on data from different domains.

5.2 Effectiveness of transfer learning

The comparison of baseline results, and results after fine tuning was applied can be seen in the plots. All of the results are shown in Table 3. We notice an increase in performance across all 3 models across all 4 evaluated datasets when the LOO approach is used, and performance more than double from 30% to 70% for the Amazon dataset in Fig. 2 and from 40% to 80% for the SST-2 subset in Fig. 5, and significant increases for the IMDB dataset in Fig. 4 and Financial Phrasebank dataset in Fig. 3. This clearly shows that fine tuning works well in general, and even when fine tuned and evaluated on datasets in different domains (Amazon vs Financial).

In the case of the adversarial dataset, all models are expected to perform worse as the dataset is designed to trick the classifier. This is the case shown on both the SST-2 and IMDB dataset, but the models seem to do significantly better on the Financial Phrasebank dataset, and slightly better on the Amazon dataset. This suggests that models fine tuned on the Amazon, IMDB, and SST-2 datasets (evaluated on Finance) are more robust than other combinations, performing well even on data outside of their domains. This paper however acknowledges that as the adversarial dataset is self-constructed, there can be unforseen biases during evaluation which leads to unexpected results.

In both Fig. 2 and 4, where DistilBERT and BERT performed worse on the adversarial datasets, RoBERTa stands out as the only model performing better. This suggest one of the benefits of having an optimised model is that is generalizes relatively better after fine tuning as opposed to other models, and if the size of the datasets for fine tuning were bigger, we could perhaps see RoBERTa performing decisively better than the other two in evaluation of the adversarial dataset.

Table 3 shows the weighted F1-scores of all the experiments we have done across all the models, and Fig. 4 represents the mean and standard deviation of the evaluations. We notice a single combination of F, A, S, that is fine tuning done on Financial Phrasebank, Amazon Polarity, and then SST-2 subset, doing the best across almost all evaluations on the IMDB dataset. This shows there is a single combination and order of fine tuning that significantly improves the performance of models. One possible explanation is that the IMDB and Amazon datsets are more semantically similar to SST-2 than the Financial Phrasebank, and fine tuning the model on I and A last provides the models with weights that are more optimal; This is also shown in the combination and $F, S, A \rightarrow I$ which had the second best performance in the group of evaluations on IMDB, with DistilBERT achieving 90.4% and BERT achieving 89.6%. The performance of these models of the combination F, A, S also performed well on the adversarial dataset.

	Leave-One-Out			Adversarial		
Methods	DistilBERT	RoBERTa	BERT	DistilBERT	RoBERTa	BERT
$A, F, I \rightarrow S$	0.850	0.878	0.854	0.850	0.860	0.818
$A, I, F \rightarrow S$	0.863	0.892	0.879	0.779	0.841	0.818
$I, A, F \rightarrow S$	0.845	0.344	0.860	0.779	0.402	0.800
$I, F, A \rightarrow S$	0.852	0.891	0.875	0.801	0.840	0.781
$F, A, I \rightarrow S$	0.847	0.344	0.856	0.779	0.402	0.800
$F, I, A \rightarrow S$	0.854	0.881	0.868	0.780	0.820	0.780
$A, I, S \rightarrow F$	0.736	0.528	0.560	**0.920**	0.402	0.841
$A, S, I \rightarrow F$	0.703	0.528	0.669	0.800	0.402	0.860
$I, A, S \rightarrow F$	0.624	0.293	0.669	0.860	0.640	0.940
$I, S, A \rightarrow F$	0.660	0.528	0.583	0.820	0.402	0.800
$S, A, I \rightarrow F$	0.722	0.789	0.700	0.841	0.880	0.879
$S, I, A \rightarrow F$	0.685	0.528	0.626	0.820	0.402	0.820
$A, F, S \rightarrow I$	0.888	0.336	0.886	0.900	0.402	0.861
$A, S, F \rightarrow I$	0.867	0.336	0.823	0.879	0.402	0.879
$F, A, S \rightarrow I$	**0.905**	**0.929**	**0.904**	0.860	**0.880**	**0.960**
$F, S, A \rightarrow I$	0.904	0.336	0.896	0.860	0.402	0.800
$S, A, F \rightarrow I$	0.887	0.336	0.894	0.780	0.402	0.819
$S, F, A \rightarrow I$	0.898	0.336	0.896	0.820	0.402	0.821
$I, F, S \rightarrow A$	0.871	0.344	0.864	0.899	0.402	0.840
$I, S, F \rightarrow A$	0.878	0.458	0.860	0.860	0.483	0.840
$F, I, S \rightarrow A$	0.877	0.344	0.876	0.841	0.402	0.880
$F, S, I \rightarrow A$	0.887	0.344	0.892	0.879	0.402	0.841
$S, I, F \rightarrow A$	0.897	0.928	0.897	0.820	0.819	0.860
$S, F, I \rightarrow A$	0.891	0.344	0.880	0.799	0.402	0.734

Table 3: Weighted F1-Scores for different methods and datasets. A stands for the Amazon dataset, F for Financial Phrasebank, I for IMDB, and S for SST-2. $(A, F, I \rightarrow S)$ in the first column means that the model has been fine-tuned on Amazon, Financial Phrasebank, and IMDB, and evaluated on SST-2, using the Leave-One-Out approach. The second set of columns represent the same combination evaluated on the adversarial dataset. Highlighted in bold are the best results for each model.

This could be due to the adversarial dataset containing some entries that share a domain similarity with the IMDB dataset. The fine tuned weights on the model will correspond well to both the IMDB and adversarial datasets, thus giving the best performance in both evaluations.

5.3 Robustness of model across fine tuning

Table 4 shows the mean and standard deviation of performance of models across all evaluation datasets. It shows that DistilBERT and BERT emerges as more robust, with DistilBERT performing the best on evaluation of the IMDB, and Financial Phrasebank datasets, while BERT performs the best in the SST-2 and adversarial datasets. This suggests that models which are not optimised and are not data hungry, in fact performs better across different datasets, and by large margins.

5.4 Impact of fine tuning order

The results also show the big impact of order of fine tuning using different datasets. Comparing across LOO evaluations on the IMDB dataset, the difference in performance between the combination F, A, S and A, S, F is 3.8% for DistilBERT, 8.1% for BERT, and a significant 59.3% for RoBERTa. This result is also shown clearly in Table 4, where the standard deviation in evaluation of the Amazon dataset reached as high as $\pm 23.5\%$. This suggests a very clear importance in performing cross validation across all combinations when fine tuning a model on multiple datasets, to determine the optimal order, as the order significantly affects the variance in performance.

Evaluation Dataset	DistilBERT	RoBERTa	BERT
Amazon	0.703 (0.235)	**0.730 (0.184)**	0.703 (0.235)
IMDB	**0.892 (0.001)**	0.435 (0.221)	0.883 (0.027)
Financial Phrasebank	**0.688 (0.038)**	0.532 (0.143)	0.635 (0.050)
SST-2	0.861 (0.020)	0.803 (0.206)	**0.873 (0.016)**
adversarial	0.802 (0.136)	0.662 (0.225)	**0.804 (0.141)**

Table 4: Mean weighted F1-Score for different evaluation datasets, with the standard deviation in brackets.

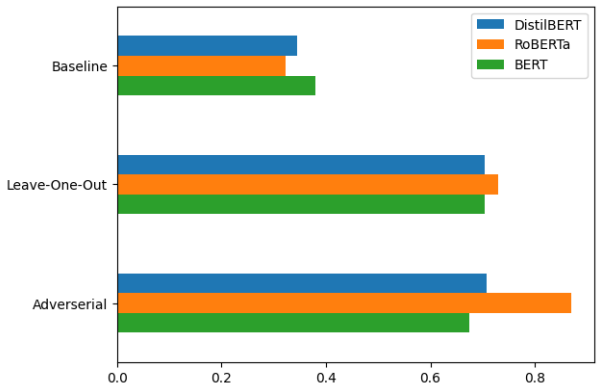

Figure 2: Comparison of weighted F1-scores between Baseline, Leave-One-Out, and adversarial Methods for the Amazon Polarity dataset.

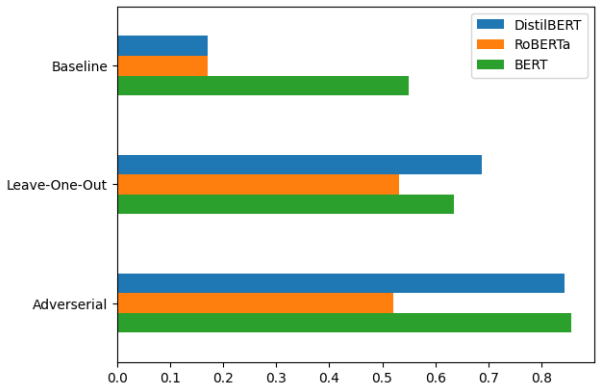

Figure 3: Comparison of weighted F1-scores between Baseline, Leave-One-Out, and adversarial Methods for the Financial Phrasebank dataset.

6 Discussion and Conclusion

In our experiments, we demonstrated differences in evaluation of datasets across different models, which suggests how the differences in model architectures could potentially lead to drastically different performances. We showed the positive efficacy of transfer learning, leading to significant improved performances when models are fine tuned over multiple datasets. The robustness of models were measured by the consistency of their weighted F1-scores across different fine tuning and evaluations, with DistilBERT and BERT being ahead in terms of generalization of training. Finally, we also made clear the significant impact of the combination of fine tuning (i.e. the order of datasets in which a model is fine tuned on), which suggests that this can be a hyperparameter for which validation is used to identify the optimal combination for the given task.

The results were mainly judged on the weighted F1-scores of the evaluations, which takes into account the proportion of classes in the datasets. While this metric is ideal for our experiments where there is class imbalance, Micro/Macro F1-scores could be used to judge a different perspective of the results, such as that of the ratio of total true positives and negatives. The PRC/AUC could also be used to judge the tradeoff between precision and recall. The purpose of this paper is to investigate if models remain robust across different datasets, and we chose one specific metric to ground and focus our findings; There could potentially be interesting insights if a different metric is sought.

For the model hyperparameters, we decided to select those that are baseline and provide sensible results for the evaluations. However, we acknowledge that there can be variance if different hyperparameters were used. For example, if RoBERTa were to use specific hyperparameters that reduces its optimisation, it could then require less data for an improved performance, and thus perform better across the experiments.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

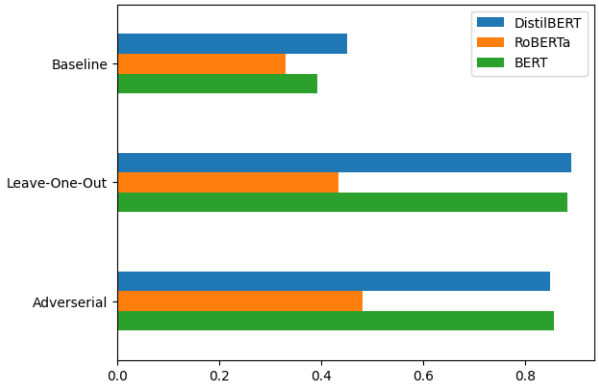

Figure 4: Comparison of weighted F1-scores between Baseline, Leave-One-Out, and adversarial Methods for the IMDB dataset.

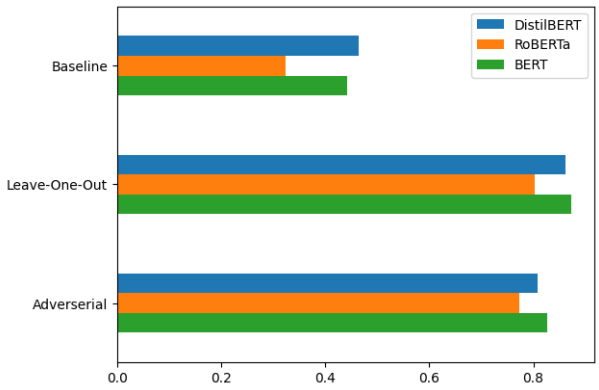

Figure 5: Comparison of weighted F1-scores between Baseline, Leave-One-Out, and adversarial Methods for the subset of the SST-2 dataset.

Hai Ha Do, Penatiyana WC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications*, 118:272–299.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware bert for cross-domain sentiment analysis. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 4019–4028.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751*.

Fatima-Ezzahra Lagrari and Youssfi Elkettani. 2021. Traditional and deep learning approaches for sentiment analysis: A survey. *Advances in Science, Technology and Engineering Systems Journal*, 6(4):1–7.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yuta Shingu, Yuki Takeuchi, Suguru Endo, Shiro Kawabata, Shohei Watabe, Tetsuro Nikuni, Hideaki Hakoshima, and Yuichiro Matsuzaki. 2021. Variational secure cloud quantum computing. *arXiv preprint arXiv:2106.15770*.

Yang Tan, Yang Li, Shao-Lun Huang, and Xiao-Ping Zhang. 2022. Transferability-guided cross-domain cross-task transfer learning. *arXiv preprint arXiv:2207.05510*.

Karl R Weiss and Taghi M Khoshgoftaar. 2017. Analysis of transfer learning performance measures. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 338–345. IEEE.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

A Additional Resources

The full adversarial dataset can be found at: `https://huggingface.co/datasets/nathandsouza10/comp0087-snlp` Table 5 presents a selection of adversarial examples from the dataset. These examples showcase a variety of linguistic techniques used to create challenging scenarios for sentiment analysis models.

IMDB Dataset The IMDB dataset can be found at: `https://huggingface.co/datasets/stanfordnlp/imdb`

Text	Label	Technique Description
"Wow, thanks for refilling my water once the entire meal. Truly exceptional service!"	0	Sarcasm - The positive expression is used sarcastically to criticize poor service.
"These instructions are crystal clear. Even a toddler could understand them."	1	Hyperbole - Exaggerates the simplicity of the instructions for emphasis.
"This movie is like watching paint dry, except paint drying might actually be more exciting."	0	Indirect Expression - The statement indirectly expresses boredom or dissatisfaction with the movie.

Table 5: Subset of Adversarial Examples Highlighting Linguistic Techniques to Challenge Sentiment Analysis

Financial Phrasebank Dataset The financial phrasebank dataset can be found at: https://huggingface.co/datasets/financial_phrasebank

Amazon Polarity Dataset The Amazon Polarity dataset can be found at: https://huggingface.co/datasets/amazon_polarity

SST2 Dataset The SST2 dataset can be found at: https://huggingface.co/datasets/stanfordnlp/sst2

GitHub Repository The GitHub training pipeline can be found at https://github.com/ali-soomro/COMP0087-SNLP