Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4235–4244
Marseille, 20-25 June 2022

Transfer Learning Methods for Domain Adaptation in Technical Logbook Datasets

Farhad Akhbardeh, Marcos Zampieri, Cecilia Ovesdotter Alm, Travis Desell
Rochester Institute of Technology
Rochester, NY, USA
{fa3019, mazgla, coagla, tjdvse}@rit.edu

Abstract

Event identification in technical logbooks poses challenges given the limited logbook data available in specific technical domains, the large set of possible classes, and logbook entries typically being in short form and non-standard technical language. Technical logbook data typically has both a domain, the field it comes from (*e.g.*, automotive), and an application, what it is used for (*e.g.*, maintenance). In order to better handle the problem of data scarcity, using a variety of technical logbook datasets, this paper investigates the benefits of using transfer learning from sources within the same domain (but different applications), from within the same application (but different domains) and from all available data. Results show that performing transfer learning within a domain provides statistically significant improvements, and in all cases but one the best performance. Interestingly, transfer learning from within the application or across the global dataset degrades results in all cases but one, which benefited from adding as much data as possible. A further analysis of the dataset similarities shows that the datasets with higher similarity scores performed better in transfer learning tasks, suggesting that this can be utilized to determine the effectiveness of adding a dataset in a transfer learning task for technical logbooks.

Keywords: Transfer Learning, Domain Adaptation, Technical Logbooks, Event Classification

1. Introduction

Estimating downtime and performing timely maintenance is a key step in reducing costs, increasing safety, and improving operational efficiency in various branches of engineering. Predictive maintenance systems use machine learning to estimate when maintenance operations should occur by using many sources of information, including historical maintenance records in the form of event logbooks (Carvalho et al., 2019).

Event logbooks often contain short text fields that describe an issue or a solution taken to address a safety problem (*e.g.*, replacing a part). They also often include issue type labels manually assigned by domain experts, making it possible to train systems to classifying maintenance issues automatically according to similarity (McArthur et al., 2018). The issue type and descriptions include domain-specific technical language, abbreviations, and non-standard orthography, which off-the-shelf NLP models are unable to process. This has led to development of domain-specific text pre-processing pipelines for logbook entries (Deléger et al., 2010; Akhbardeh et al., 2020b).

Another import challenge when processing logbook datasets is data scarcity. Logbooks are proprietary and not widely available making it difficult to train robust predictive maintenance systems that require large amounts of data, for example, when using deep neural networks. One exception is MaintNet, an open repository of predictive maintenance datasets from multiple domains, such as automotive and aviation (Akhbardeh et al., 2020a). As evidenced in Section 3, these technical logbook datasets are, nonetheless, relatively small–ranging from a few hundred instances for automotive maintenance to nearly $75,000$ instances for facility maintenance.

To address the limitation of data size, we explore the use of transfer learning for domain adaptation in technical event classification. Transfer learning techniques have been applied with great success to non-standard (*e.g.*, social media posts) text-based classification tasks such as sentiment analysis and offensive language identification (Tao and Fang, 2020; Ranasinghe and Zampieri, 2020) suggesting that these techniques can also be successfully applied to technical logbooks. To the best of our knowledge, however, transfer learning techniques have not been yet applied to technical logbook datasets and our work addresses this gap. This work contributes with a comprehensive study of domain adaptation under three conditions: (1) transfer within a domain, (2) transfer within an application, and (3) transfer over a global dataset.

We address the following research questions:

- **RQ1:** Which transfer learning approaches are better suited for classifying technical events for predictive maintenance across heterogeneous logbook datasets?
- **RQ2:** How does the level of similarity between corpora impact the performance of transfer learning approaches for technical event classification?

2. Related Work

Transfer Learning Transfer learning strategies have been applied in various NLP tasks such as sentiment

Inst.	Example Instance of Technical Logbook Entry	Technical Event Type	Domain Terms & Abbr.
1	LANDING AIRCRAFT LOST **ALTITUDE** WHILE TURNING BASE TO FINAL	SUBSTANTIAL DAMAGE	ALTITUDE
1	AIRCRAFT RIGHT **GEAR** CATCH FIRE ON **RWY** DALLAS **TX**	MINOR DAMAGE	GEAR, RWY, TX
2	**R/H FWD** UPPER **BAFF** SEAL NEEDS TO BE RESECURED	BAFFLE DAMAGE	R/H, FWD, BAFL
2	LEFT **ENG I/B** BAFFLE INTERCONNECT **ROD** BROKEN	BAFFLE LOOSE	ENG, I/B, ROD
3	ABNORMALITES NOTE **FAN BLAD** BEND OUTWARD POST FLIGHT **INSP**	CAUSED DAMAGE	FAN, BLAD,INSP
3	**ENG** PARAMETERS NORMAL, BUT NEEDS **INSP**	NO DAMAGE	ENG, INSP
4	CHECK **L/R** OUTER TIRE, AND **GAS PADDLE**	PM SERVICE	L/R, GAS, PADDLE
4	CHECK **OIL** OR TRANS LEAK **SPINNER** LIGHT ADJ **CONV** CHAIN PLOW	DRIVER REPORTED	OIL, SPINNER, CONV
4	**PLOW** DONT WORK ROAD CALL **CK BATTERY**	BREAKDOWN	PLOW, CK, BATTERY
5	**BRAKE** FAILURE OR DEFECTIVE	NON-INCAPACITATING	BRAKE
5	DISREGARDED THE **SIGNAL** OR REGISTRAR **SIGN**	UNKNOWN	SIGNAL, SIGN
6	CLEANED AROUND THE EXTERIOR OF THE **BLDG**	SERVICE	BLDG
7	PRETTY CONSISTENT **SPEEDING** ALL HOURS OF THE DAY	SPEEDING	SPEEDING
7	EXCESSIVE **SPEEDING** ALONG ARKANSAS	SPEEDING	SPEEDING
7	**CARS** TRYING TO GET TO THIS INTERSECTION ON THE REGULAR	BLOCKING CROSSWALK	CARS

Table 1: Instances of technical logbook entries by domain experts spanning aviation accident (1), aviation maintenance (2), aviation safety (3), automotive maintenance (4), automotive accident (5), facility maintenance (6), and automotive safety (7). Instances have domain-specific terminology (Terms.), abbreviations (Abbr.), and nonstandard forms. Details are in Section 3.

analysis to address a deficit of labeled data. Studies have shown that training a model on one task or dataset (the *source*) and using transfer learning methods to transfer the knowledge to another task or dataset (the *target*) can improve performance compared to the model trained only on the target task with less data (Pan and Yang, 2010). Tao and Fang (2020) proposed a transfer learning approach to overcome limited annotated data for aspect-based sentiment analysis tasks. Their analysis includes applying sentiment datasets from different domains to evaluate performance on sentence-level multi-label classification using the XLNet (Yang et al., 2019) and BERT (Devlin et al., 2019) models, improving over baseline methods.

Terechshenko et al. (2020) took advantage of transfer learning in political data analysis to overcome a limited dataset. They employed the XLNet model to transfer learned knowledge to political science texts. Their experiment identified improvement on using a small source of a labeled dataset for transfer learning.

Transfer learning has also been used in healthcare to address the bottleneck of large labeled datasets and enable generalization capability. Romanov and Shivade (2018) proposed a transfer learning technique to utilize the open-source Stanford Natural Language Inference dataset and medical terminologies in expert-annotated clinical data. They experimented with sequential inference using an LSTM in multiple layers. Their approach improved over previously reported methods on Natural Language Inference benchmarks. Dirkson and Verberne (2019) proposed a transfer learning method to Twitter data associated with health to classify drug effects. They utilized a recurrent neural network architecture by Flair (Akbik et al., 2019) having 512 hidden layers, yielding performance improvements over a baseline support vector classifier (SVC). Their finding was further that it can be beneficial to apply various domain-specific domain adaptation strategies.

Domain Adaptation Employing domain adaptation to transfer learned knowledge of a source domain and improve performance in a target domain has also shown success in NLP problems. Axelrod et al. (2011) investigated domain adaptation in statistical machine translation by utilizing instances from a general domain translation corpus of English and Chinese. They employed Moore and Lewis (2010)'s domain-specific models. The approach was successful with just a small subset of in-domain data. Heilman and Madnani (2013) utilized Daumé III (2007)'s domain adaptation in automatic short answer scoring and combined n-gram features and corpus similarity measures in the educational domain. Their results had high accuracy which approached human scores on the Beetle dataset that consists of student short answers to numerous questions.

Furthermore, several cross-domain adaptation approaches have been developed to leverage the knowledge learned in one domain to another. Peng and Dredze (2017) studied transferring multi-task learning representations for sequence tagging using news and social media texts. They proposed a multi-task framework based on the BiLSTM model, which was capable of sharing the learner representation across tasks or domain datasets. The proposed framework achieved higher accuracy when applied to social media. El Mekki et al. (2021) proposed an unsupervised domain adaption approach utilizing pre-trained language models for cross-dialect sentiment analysis. They examined incorporating the Arabic dialects' fine- and coarse-grained taxonomies. In comparison to the zero-shot transfer using the BERT model, their approach showed roughly 20% performance improvement using the cross-dialect domain adaptation approach.

The many applications of transfer learning in NLP confirm that transfer learning approaches are a promising

strategy to overcome data scarcity. However, the use of transfer learning to predictive maintenance datasets has not yet been explored. Our work fills this important gap providing empirical evidence of the feasibility of these methods when applied to technical logbook data.

3. Technical Event Datasets

In this work we used 7 technical logbook datasets in English, presented in Table 2. These datasets are from three domains of aviation, automotive, and facilities, available at MaintNet (Akhbardeh et al., 2020a). Three datasets are from the aviation domain: aviation maintenance (*Avi-Main*) featuring 7 years of historical maintenance logbook, aviation accidents (*Avi-Acc*) contains 4 years of aviation accident and reported damages, and aviation safety (*Avi-Safe*) contains 11 years of aviation safety. Three other datasets are from the automotive domain: automotive maintenance (*Auto-Main*) contains a single year report of car maintenance, the automotive accident (*Auto-Acc*) contains 12 years reporting about car accidents and crashes, and automotive safety (*Auto-Safe*) contains 4 years of driver's noted hazards and incidents on the roadway. And finally, we also used a single dataset of facility maintenance (*Faci-Main*) that contains 6 years of logbook reports collected for building maintenance.

Dataset Description The instances in the technical logbook dataset consist of a compact and brief summary of a problem that occurred. This problem description is mainly composed of domain-specific abbreviations, vocabularies, and terminologies. As shown in Table 1, these instances are single sentences that usually contain short text. The instances such as "*eng light on, remoev hyd lines, leak note*" in aviation maintenance data (*Avi-Main*) or "*ckd fire ext throughout bldg*" in facility maintenance (*Faci-Main*) contain domain-specific abbreviations (*eng, hyd, bldg*), or misspelling (*remoev*) that briefly forms the description regarding specific event type. Further, technical logbook datasets in these different domains contain terms and abbreviations that are similar or identical but with different meanings when appearing in a different domain, and are non-standard to typical pre-processing pipeline packages. For example, in the instance "*while in fl, after performing a few high power man*" where *fl* refers to the *flight level* and *man* refers to *manual*, rather than to their typical expansion (*Florida*) or lexical sense (*male individual*). Further highlighting the non-standard lexicon of these datasets, Figure 1 provides the top 10 most frequent words in aviation domain datasets.

Domain	Dataset	Instance	Class	Code
Aviation	Maintenance	6,169	39	*Avi-Main*
	Accident	4,130	5	*Avi-Acc*
	Safety	17,718	2	*Avi-Safe*
Automotive	Maintenance	617	5	*Auto-Main*
	Accident	52,707	3	*Auto-Acc*
	Safety	4,824	17	*Auto-Safe*
Facility	Maintenance	74,360	70	*Faci-Main*

Table 2: The number of different instances with various size in each technical domain dataset.

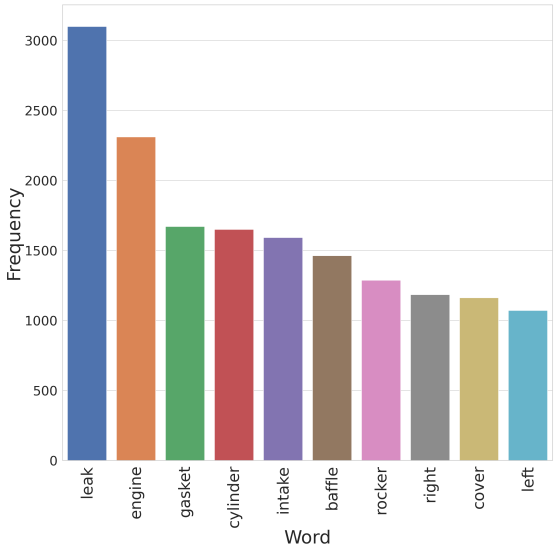

Figure 1: Top 10 frequency of words used in the aviation maintenance (Avi-Main) datasets representing the nature of such technical logbook data.

Dataset Challenges In addition to these previously discussed challenges, there are additional key challenges related to technical logbook data that make it difficult for off-the-shelf NLP pipelines to handle. In a general language corpus (*e.g.*, news text, Wikipedia text), the instance usually follows standard formatting and structure that current NLP models (*e.g.*, pre-trained language models) can process properly. However, the written description in the technical logbook lacks such a standardized structure due to the different writing formats that domain experts use while describing the observed issue during an inspection. Furthermore, the aforementioned non-standard format of technical logbooks poses various challenges to the machine learning model. The challenges that need to be considered are:

1) Utilizing various domain-specific abbreviations and acronyms that might be specific to each domain (*e.g.*, *AGL – Above Ground Level*) and dropping or substituting any character can alter the meaning (*e.g.*, *AGL* to *AL – Approach Lights* or *ALS – Approach Lighting System*); 2) Using uncommon syntax and parts of speech sequences (*e.g.*, *tires, lights testing showed multiple issues*) or contractions (*e.g.*, *needn't – need not*); 3) Using misspelling or dropped words in the description which can be inaccurately seen as abbreviations (*e.g.*, *fast* where dropped to: *fas – final approach segment*); and 4) Using problem descriptions of varying length that can consist of a few tokens that describe the same issue (*e.g.*, *engine failed, engine not working properly*).

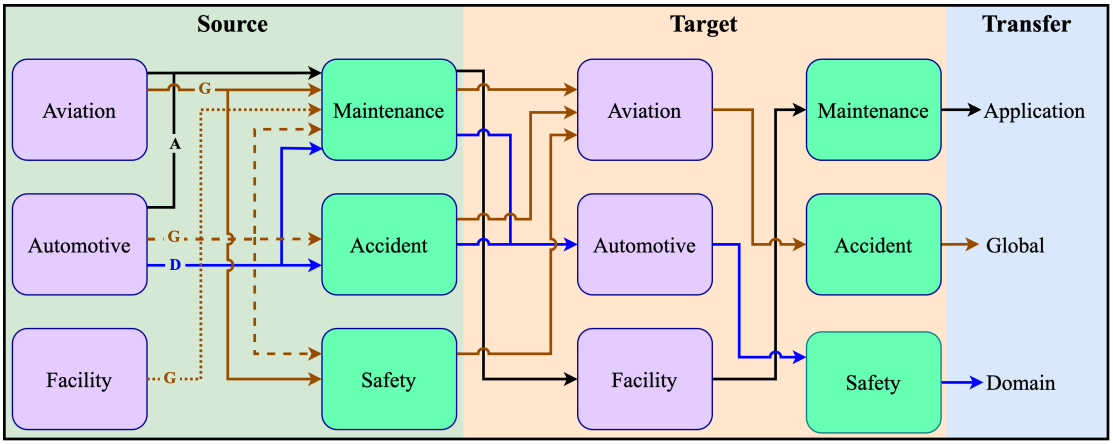

Figure 2: Process of transfer learning methods for technical logbooks by representing the three various approaches of transferring within the domain, transferring within an application, and transferring over the global dataset. The black color (A) represents to application, blue (D) represents the domain, and brown (G) represents the global transferring method. As an example of transferring within an application, we train the model on aviation maintenance (*Avi-Main*) and automotive maintenance (*Auto-Main*) source datasets and then take the trained model and transfer it to the facility maintenance (*Faci-Main*) data as the target dataset to perform further training and classification.

4. Methods and Models

To quantify the impact and success of transfer learning on the technical datasets, we first performed event identification using a classifier on a single source domain dataset to serve as our baseline results, and then we utilized the same model to perform transfer learning with the following proposed strategies. In this experiment, both the source and target datasets are labeled which enables the use of supervised learning methods.

4.1. Transfer Learning of Technical Logbooks

Text classification using a modest, small technical dataset, such as the automotive maintenance data, can limit the model's generalization capacity and performance. One potential solution could be to utilize a data augmentation approach to increase the dataset size by generating synthetic data. However, domain-specific datasets where each domain captures domain-specific lexical semantics – the case for technical logbooks as illustrated previously – prevents the use of techniques such as domain-discriminative data selection applied to the smaller domain data class (Ma et al., 2019). Furthermore, these technical datasets are highly imbalanced.

Therefore, we studied four different methods of transfer learning (domain adaptation) using the seven domain-specific datasets and analyzed their effects on the performance of technical event classification on the target datasets: (1) a baseline strategy which simply trains the model on a single dataset, and then also strategies that (2) transfer to a dataset from other sources within the domain (but different applications), (3) transfer to a dataset from sources with the same application (but different domains), and (4) transfer from all other sources in the global dataset. Figure 2 provides an overview of the process of transferring from the source to the target dataset utilizing three transfer learning methods in this work.

Transferring within a Domain Transferring a learned model within a domain dataset can benefit the target domain dataset by utilizing knowledge learned from various domain datasets, as these corpora should have similar vocabularies. In this approach, we train the model on selected datasets within the domain, and then transfer the model to a different target dataset, where we continue to train the model to perform event classification. As an example, we can train the model on the aviation maintenance (*Avi-Main*) and aviation safety (*Avi-Safe*) source datasets and then take the trained model and transfer it to the aviation accident (*Avi-Acc*) data as the target dataset, to perform further training and classification.

Transferring within an Application While datasets within a domain stand to share similarities between their corpora, datasets that share an application may also stand to benefit in a similar manner, however with a different set of potential vocabulary. This strategy evaluates the impact of the shared knowledge within an application, where the model is first trained on source datasets from other domains which share the same application, and then transferred to the target dataset. For example, the model can be first trained on the source dataset of aviation accidents (*Avi-Acc*) and then transfer to the target domain of automotive accidents (*Auto-Acc*) for further training and classification.

Transferring over the Global Dataset Finally, to provide another option to address the potential that transfer learning may be improving performance sim-

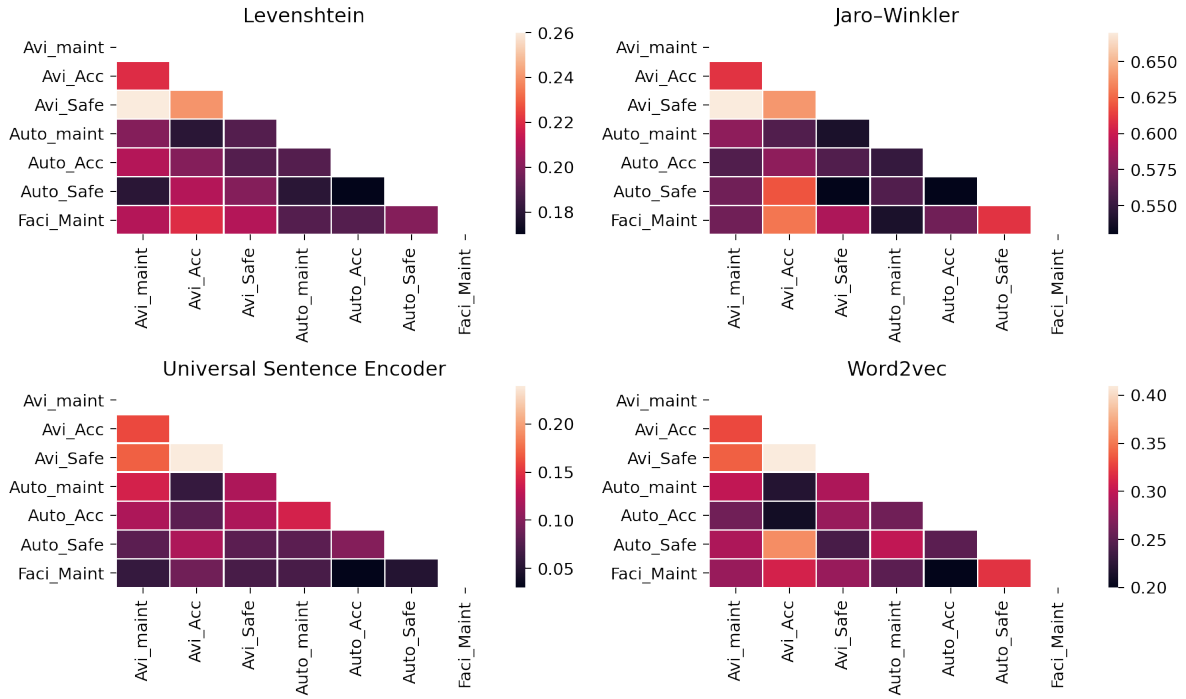

Figure 3: Heat maps of similarity scores for 4 algorithm including Levenshtein, Jaro-Winkler, Universal Sentence Encoder, and Word2vec applied on 500 random instances from each domain. The various color types shown in these figures describe when the similarities between a pair of datasets are lower or higher where the higher value (with lighter color) defines higher similarity, and lower value (with darker color) define low similarity.

ply because the model had more data to train on (Melamud et al., 2019), a global transfer strategy is investigated. In this strategy, given the seven datasets available, we considered every single dataset as a target, and then initially train the model on every other dataset available.

4.2. Model Architecture and Training

For the event identification task, we considered a supervised machine learning method to classify the issue type (*e.g.*, cylinder damage, intake gasket leak). As mentioned above, the event description in a dataset contains short text and has a single event category. The machine learning model used in this study is a convolutional neural network (CNN) (Kim, 2014) model that has shown success in several NLP tasks such as question classification (Shen et al., 2018) or sentiment analysis (Wang et al., 2016), and that further is capable of providing suitable performance while applying it on various sequence types. Furthermore, we also evaluated it using the pre-trained ALBERT (Lan et al., 2020) for English and we fine-tuned the model on the downstream task of event identification. We used the ALBERT transformer model in this work as has previously been shown to achieve high performance in various NLP tasks and benchmarks using less parameters than other transformers models, and that it also benefits from a cross-layer parameter sharing property (Kung et al., 2020).

We trained the CNN architecture proposed by Kim (2014) which consists of 100 one-dimensional convolutional filters with the size of multi n-gram lengths followed by ReLU activation, dropout layer, max-pooling layer with the size of 2 by the length of the input sequence and followed by a fully connected dense layer and output layer set to class size with SoftMax activation function.

As discussed in Section 4.1, the technical dataset classes are highly imbalanced in general. These characteristics can cause the classification model to overgeneralize the majority class. To address this issue, we utilized the feedback loop strategy initially proposed by Bowley et al. (2019) in computer vision and which has since been successfully adapted to the NLP domain (Akhbardeh et al., 2021). This approach not only mitigates the problem of the classifier preferring larger majority classes but also adjusts the training data such that the model keeps training on the worst performing training instances.

4.3. Datasets and Baselines

To address the issues with technical logbook datasets noted in Section 3, we utilized the text pre-processing pipeline developed by Akhbardeh et al. (2020a) which is capable of domain-based abbreviation expansion, noise entity removal, lexical normalization, dictionary-based standardization, part of speech tagging, and domain-specific lemmatization. The dataset is divided

into 80% for training and 20% for testing, and 100 dimensional word embeddings (Mikolov et al., 2013) were used for feature extraction.

Baseline The baseline strategy for training the models consists of training the model on a single dataset (*e.g.*, *Auto-Safe*), and then perform the classification task on the 7 datasets. The baseline strategy does not contain any transfer learning approaches and solely uses the source dataset.

Technical Logbook Similarity Technical logbooks contain different instance sizes and token sizes, however as described in Section 3, they contain usually short instances, as well as specific words, terms, and abbreviations, where they have less ambiguity. These instances in the logbook dataset also share similar or dissimilar characteristics that would be used in specific terms or abbreviations. Each domain dataset has similar terminologies that are shared with other domains (*e.g., ft - feet*), however, some can share similar abbreviations by different semantics (*a/c* - in aviation domain: aircraft, in automotive and facility domain: air-conditioner). These shared similarities and features in the datasets could bring useful information when training the model on a specific dataset(s) and transferring the learned knowledge between within domains or dataset(s).

Instances in datasets such as those in the aviation domain, contain descriptions regarding problem types that are semantically similar to other domain's dataset such as the automotive domain (*e.g.*, *engine not start*, *cylinder compression issue*). Evaluation of instance similarity within the logbook dataset can help to interpret how data are semantically similar in either a word or sentence level. This can be done by measuring the inter-corpus similarity and identifying corpus homogeneity (Cavaglià, 2002). We experimented with applying four similarity measures including Levenshtein (Konstantinidis, 2007), Jaro-Winkler (Wang et al., 2017), Universal Sentence Encoder (Cer et al., 2018), Gensim Word2vec (Rehurek and Sojka, 2011) to compare and extract key relations between instances in the dataset. In both the Universal Sentence Encoder and Word2vec model, we utilized the cosine similarity (Manning et al., 2008) for computations.

The corpus similarity experiment in this study has been done using random sets of 500 instances from each of seven dataset and we calculated the similarity measurements between instances in an inter-document form. This means every instance from a selected dataset was compared to the instances in the other remaining selected datasets to compute the distance. Figure 3 shows the findings of these analyses in heat maps and further discussion regarding the relationship between corpus similarity and domain adaptation is provided in Section 5.

5. Results

This section provides a performance analysis of transfer learning between varying dataset types using the previously described CNN and ALBERT (transformer-based) model. Table 3 presents an evaluation of training only on the source dataset to transferring models trained on other datasets (either within the domain, within the application, or all other datasets) to that source dataset.

Experimental Settings In the experimental process, we used the coarse to fine learning approach for optimizing parameters and hyperparameters (Lee et al., 2018). Hyperparameters for training set by investigating batch sizes of 32, 64, and 128 (Masters and Luschi, 2018), with an initial learning rate of 0.01 for CNN and 1e-5 for fine-tuning ALBERT model. Further, dropout regularization ranging from 0.2 to 0.4, and ReLU and SoftMax as an activation function (Nair and Hinton, 2010), and Adam optimizer (Kingma and Ba, 2015), and categorical cross-entropy (Zhang and Sabuncu, 2018) for the loss function were selected. Based on the experiment and model performance, dropout regulation with a rate of 0.2 and batch size of 32 were selected for the training.

Experimental Design First, the CNN and ALBERT (fine-tuned) model was trained 10 times on each source dataset, with the *baseline* (source) column reporting the average precision (P), recall (R) and F1 scores of those runs. Following this, 10 CNN and ALBERT (fine-tuned) models were trained on the other domains, and then each of those 10 models were transferred to the source dataset for further training (layer freezing was not used). These results are reported in the *domain*, *application* and *global* columns. Additionally, Mann-Whitney U-Tests of statistical significance were performed comparing the populations of final losses across the 10 repeated experiments of the source data, to the final losses of the 10 repeated experiments for each of the transferred runs, which are reported in the *S* columns. Based on these experiments, we observed performance improvements for each dataset.

6. Discussion

All the datasets in the aviation domain had improved performance when being transferred to from within the domain with statistical significance, while their performance was degraded when transfer learning was performed across applications and from the global dataset. Similar results were found in the automotive domain, where the model performance improved across all datasets when using the within-domain transfer learning approach, however, interestingly, the automotive accident (*Auto-Acc*) dataset also achieved better model performance while transferring over the application and global datasets, with the best performance coming from the global dataset. For the other two datasets, within-domain transfer learning found the best results,

Dataset	Model	Baseline (%)			Domain (%)				Application (%)				Global (%)			
		P	R	F1	P	R	F1	S	P	R	F1	S	P	R	F1	S
Avi-Main	CNN	0.91	0.89	0.89	0.92	0.89	***0.90***	**0.0003**	0.89	0.88	0.88	0.1510	0.87	0.87	0.87	**0.0472**
	ALBERT	0.89	0.87	0.88	0.90	0.91	***0.90***	**0.0014**	0.88	0.87	0.87	0.1918	0.86	0.85	0.85	**0.0040**
Avi-Safe	CNN	0.88	0.85	0.86	0.89	0.87	***0.88***	**0.0336**	0.89	0.82	0.85	**0.0066**	0.88	0.82	0.85	**0.0028**
	ALBERT	0.87	0.84	0.85	0.88	0.86	**0.87**	**0.0124**	0.86	0.82	0.84	**0.0092**	0.85	0.85	0.84	**0.0163**
Avi-Acc	CNN	0.49	0.51	0.49	0.50	0.51	***0.50***	**0.0056**	0.48	0.51	0.48	**0.0293**	0.48	0.50	0.49	0.2982
	ALBERT	0.44	0.48	0.46	0.45	0.50	**0.47**	**0.0094**	0.42	0.48	0.45	**0.0187**	0.41	0.47	0.44	**0.0170**
Auto-Main	CNN	0.64	0.70	0.67	0.67	0.71	***0.69***	**0.0344**	0.64	0.68	0.66	**0.0246**	0.65	0.68	0.67	0.4548
	ALBERT	0.59	0.64	0.62	0.61	0.67	**0.64**	**0.0077**	0.58	0.61	0.60	**0.0170**	0.59	0.62	0.60	**0.0169**
Auto-Safe	CNN	0.50	0.45	0.46	0.53	0.49	***0.50***	**0.0002**	0.42	0.39	0.40	**0.0011**	0.44	0.40	0.41	**0.0171**
	ALBERT	0.48	0.46	0.46	0.50	0.48	**0.48**	**0.0026**	0.46	0.44	0.44	**0.0104**	0.47	0.42	0.43	**0.0029**
Auto-Acc	CNN	0.48	0.67	0.49	0.47	0.69	**0.50**	**0.0242**	0.47	0.68	**0.50**	**0.0372**	0.49	0.69	***0.52***	**0.0084**
	ALBERT	0.45	0.65	0.47	0.45	0.68	0.47	0.0513	0.46	0.67	**0.48**	**0.0291**	0.48	0.67	**0.48**	**0.0285**
Faci-Main	CNN	0.5	0.70	**0.56**	-	-	-	-	0.47	0.66	0.51	**0.0001**	0.45	0.65	0.49	**0.0001**
	ALBERT	0.55	0.69	***0.57***	-	-	-	-	0.51	0.67	0.54	**0.0187**	0.49	0.68	0.53	**0.0016**

Table 3: A performance comparison of the various transfer learning experiments. The average of the final models' performance across 10 repeated experiments is shown as precision (P), recall (R), F1 score, as well as statistical significance (S) using the Mann-Whitney U test. Results which outperform only training on the source dataset are in bold, and the best for a dataset are in bold and italics. Experiments which showed statistical significance with a p value of 0.05 are also in bold.

also with statistical significance (which answers research question 1). For the facilities maintenance dataset (*Faci-Main*), within-domain transfer learning was not possible as there were no other datasets in the domain, and similar to the aviation dataset, transferring from the application and global datasets also reduced performance with statistical significance.

The transfer learning results provide some interesting insights into how transfer learning can be performed across varying technical logbook datasets. While in all cases, keeping the training data within a domain provided a statistically significant improvement, in the automotive accident dataset, utilizing all other training data provided the overall best results. This suggests that while for the majority of our datasets, adding in additional training data from outside of the domain only served to confuse the models, but that this is not always the case. Some datasets may still benefit from simply having more training data due to the nature of the classification problem, or perhaps due to a wider variety of tokens allowing for more dataset similarity.

Furthermore, to address research question 2, we examined similarity measurement techniques to identify the key relationships in the aviation, automotive, and facilities domains, as well as investigating the similarity of these corpora that might lead to lowering or improving the performance of event classification model. For this reason, we applied the Levenshtein and Jaro-Winkler similarity methods to compare the similarity of these corpora. However, to further extract the key attributes between these datasets, we employed the Universal Sentence Encoder and Word2vec model to semantically evaluate the instances based on their semantic meanings. Based on the outcomes, we noticed the high inter-corpus similarity for within the aviation safety (*Avi-Safe*) and aviation maintenance (*Avi-Main*) datasets. The reason for this high inter-corpus similarity score could be the common domain terms and abbreviations that have been used in these datasets for instance "*eng was shut down and noticed slight vibration*" and "*right eng vibration with increasing power*" where the domain abbreviated word "*eng*" appeared in both aviation safety and aviation maintenance dataset respectively. Furthermore, Universal Sentence Encoder provided outcome more relates to the performance of transfer learning compared to the other methods.

7. Conclusion and Future Work

This work compared transfer learning approaches for domain adaptation for event classification in logbook datasets. We acquired seven logbook datasets from three technical domains containing short instances with non-standard grammar and spelling, and many abbreviations. We evaluated three domain adaption methods including (1) transferring within the domain, (2) transferring within the application, and (3) transferring over the global dataset compared to the baseline approach of training classification model on the single (source) domain dataset. Our results indicate that transferring within the domain dataset delivers the best performance across both CNN and ALBERT (transformer-based) models. Finally, we applied corpus similarity techniques to investigate shared characteristics among these technical datasets, using Levenshtein, Jaro-Winkler, Universal Sentence Encoder and the Gensim Word2vec models.

In future work, we will explore other methods to cope with data scarcity. This includes data augmentation techniques (Feng et al., 2021) (including rule-based and model-based), as well as zero-shot and few-shot learning classification approaches. Finally, we would also like to explore transformer-based cross-lingual methods (Ranasinghe and Zampieri, 2020) to transfer information from the English datasets available in MaintNet to low-resource languages in which there are even less predictive maintenance datasets available.

8. Bibliographical References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 54–59, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Akhbardeh, F., Desell, T., and Zampieri, M. (2020a). MaintNet: A Collaborative Open-source Library for Predictive Maintenance Language Resources. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 7–11, Barcelona, Spain, December. International Committee on Computational Linguistics (ICCL).

Akhbardeh, F., Desell, T., and Zampieri, M. (2020b). NLP Tools for Predictive Maintenance Records in MaintNet. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 26–32, Suzhou, China, December. Association for Computational Linguistics.

Akhbardeh, F., Alm, C. O., Zampieri, M., and Desell, T. (2021). Handling Extreme Class Imbalance in Technical Logbook Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4034–4045, Online, August. Association for Computational Linguistics.

Axelrod, A., He, X., and Gao, J. (2011). Domain Adaptation via Pseudo In-Domain Data Selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Bowley, C., Mattingly, M., Barnas, A., Ellis-Felege, S., and Desell, T. (2019). An analysis of altitude, citizen science and a convolutional neural network feedback loop on object detection in Unmanned Aerial Systems. *Journal of Computational Science*, 34:102 – 116.

Carvalho, T. P., Soares, F., Vita, R., da Piedade Francisco, R., Basto, J. P., and Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers and Industrial Engineering*, 137:106024.

Cavaglià, G. (2002). Measuring corpus homogeneity using a range of measures for inter-document distance. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. (2018). Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 169–174, Brussels, Belgium, November. Association for Computational Linguistics.

Daumé III, H. (2007). Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Deléger, L., Grouin, C., and Zweigenbaum, P. (2010). Extracting medical information from narrative patient records: the case of medication-related information. *Journal of the American Medical Informatics Association*, 17:555–8.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dirkson, A. and Verberne, S. (2019). Transfer Learning for Health-related Twitter Data. In *Proceedings of the Fourth Social Media Mining for Health Applications Workshop & Shared Task*, pages 89–92, Florence, Italy, August. Association for Computational Linguistics.

El Mekki, A., El Mahdaouy, A., Berrada, I., and Khoumsi, A. (2021). Domain Adaptation for Arabic Cross-Domain and Cross-Dialect Sentiment Analysis from Contextualized Word Embedding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2824–2837, Online, June. Association for Computational Linguistics.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E. (2021). A Survey of Data Augmentation Approaches for NLP. *arXiv preprint arXiv:2105.03075*.

Heilman, M. and Madnani, N. (2013). ETS: Domain Adaptation and Stacking for Short Answer Scoring. In *Second Joint Conference on Lexical and Computational Semantics: Proceedings of the Seventh International Workshop on Semantic Evalua-*

tion (SemEval 2013), pages 275–279, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Konstantinidis, S. (2007). Computing the edit distance of a regular language. *Information and Computation*, 205(9):1307–1316.

Kung, P.-N., Yang, T.-H., Chen, Y.-C., Yin, S.-S., and Chen, Y.-N. (2020). Zero-Shot Rationalization by Multi-Task Transfer Learning from Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2187–2197, Online, November. Association for Computational Linguistics.

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. In *International Conference on Learning Representations*.

Lee, K., He, L., and Zettlemoyer, L. (2018). Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 687–692, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ma, X., Xu, P., Wang, Z., Nallapati, R., and Xiang, B. (2019). Domain Adaptation with BERT-based Domain Classification and Data Selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 76–83, Hong Kong, China, November. Association for Computational Linguistics.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*, volume 39. Cambridge University Press, USA.

Masters, D. and Luschi, C. (2018). Revisiting Small Batch Training for Deep Neural Networks. *ArXiv*, abs/1804.07612.

McArthur, J., Shahbazi, N., Fok, R., Raghubar, C., Bortoluzzi, B., and An, A. (2018). Machine learning and BIM visualization for maintenance issue classification and enhanced data collection. *Advanced Engineering Informatics*, 38:101 – 112.

Melamud, O., Bornea, M., and Barker, K. (2019). Combining Unsupervised Pre-training and Annotator Rationales to Improve Low-shot Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3884–3893, Hong Kong, China, November. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations*.

Moore, R. C. and Lewis, W. (2010). Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July. Association for Computational Linguistics.

Nair, V. and Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel. International Conference on Machine Learning.

Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Peng, N. and Dredze, M. (2017). Multi-task Domain Adaptation for Sequence Tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 91–100, Vancouver, Canada, August. Association for Computational Linguistics.

Ranasinghe, T. and Zampieri, M. (2020). Multilingual Offensive Language Identification with Cross-lingual Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844. Association for Computational Linguistics, November.

Rehurek, R. and Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Romanov, A. and Shivade, C. (2018). Lessons from Natural Language Inference in the Clinical Domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium, October-November. Association for Computational Linguistics.

Shen, D., Min, M. R., Li, Y., and Carin, L. (2018). Learning Context-Sensitive Convolutional Filters for Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1839–1848, Brussels, Belgium, October. Association for Computational Linguistics.

Tao, J. and Fang, X. (2020). Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data*, 7:1, 01.

Terechshenko, Z., Linder, F., Padmakumar, V., Liu, M., Nagler, J., Tucker, J. A., and Bonneau, R. (2020). A comparison of methods in political science text classification: Transfer learning language models for politics. *Available at SSRN*.

Wang, J., Yu, L.-C., Lai, K. R., and Zhang, X. (2016).

Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 225–230, Berlin, Germany, August. Association for Computational Linguistics.

Wang, Y., Qin, J., and Wang, W. (2017). Efficient Approximate Entity Matching Using Jaro-Winkler Distance. In *Web Information Systems Engineering - WISE 2017*, pages 231–239.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhang, Z. and Sabuncu, M. R. (2018). Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8792–8802, Red Hook, NY, USA. Curran Associates Inc.