Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Doctoral Consortium Track

Transferability and Stability of Learning With Limited Labelled Data in Multilingual Text Document Classification

Branislav Pecher[1,2*]

[1]Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic
[2]Kempelen Institute of Intelligent Technologies, Mlynske nivy 5, Bratislava, Slovakia
branislav.pecher@kinit.sk

Abstract

We focus on learning with limited labelled data (especially meta-learning) in conjunction with so-far under-researched multilingual textual document classification. The core principle in such learning is to achieve transferability of learned knowledge to new datasets and tasks. Currently, factors influencing the success of transfer remain mostly unclear. Their identification from experiments is challenging due to small amounts of labels making results considerably unstable. When instability of the investigated models is not explicitly taken into consideration (as is common in existing benchmarking studies), it may result in randomness possibly even invalidating the findings. We want to remedy this by in-depth exploration of factors that influence the stability and the transferability of learning with limited labelled data in multilingual textual documents classification, such as misinformation detection.

1 Introduction and Related Work

Learning with limited labelled data, such as meta-learning, transfer learning or weakly supervised learning, is used to achieve high performance when the available labels are lacking. Applying these approaches to domains characterized by a lack of labels spread across various languages and tasks can be challenging. Achieving good transferability of learned knowledge to new datasets and tasks is vital in this setting. To do this effectively, we need to investigate what factors influence the ability to transfer knowledge to new tasks. These factors are currently mostly unknown and the negative transfer still remains an open problem [Hospedales *et al.*, 2022].

Majority of current works implicitly assume that the success of the transfer is closely tied to the similarity between data and tasks. To improve the transferability, this similarity is measured using similarity metrics or task embeddings in the models and used as weighting parameter. Benchmark datasets, many with predefined tasks splits, with very similar tasks, are used to compare the approaches. Best to our knowledge, no studies provide an in-depth analysis of the factors that affect the transferability. Only two works analyse the similarity more closely and argue for further investigation in the future, as the selection of training tasks influence the transfer in significant manner [Ye *et al.*, 2021]. Another limitation of the current state of knowledge is the focus on image data, making multilingual text classification under-researched.

A significant drawback of these studies is that they do not explicitly take into consideration their training instability. The limited availability of labels makes the training process more unstable and prone to effects of uncontrolled randomness. Even small changes in data and parameters may lead to massive changes in performance, such as changing order of samples leading from state-of-the-art results to simple random predictions [Lu *et al.*, 2021]. The significant effect of randomness may lead to the results being invalidated, when the randomness is not systematically considered. Nevertheless, the majority of existing studies perform benchmarking using a predetermined split of data and report possibly biased results from a single run. Even though some studies try to control the randomness by using more splits and doing repeated runs, no in-depth study of randomness factors exists in the context of learning with limited labelled data. The effects of randomness on the replicability are considered in few studies from other areas, such as typical deep learning. Only fraction of those, such as [Boquet *et al.*, 2019], take into consideration the interactions between the factors, instead of focusing on each factor independently.

We want to remedy these shortcomings by performing an in-depth exploration of factors that influence the stability and the transferability of learning with limited labelled data in multilingual textual document classification tasks. Our expected contributions are as follows:

1. We focus on utilizing the approaches for learning with limited labelled data in the under-researched area of multilingual text document classification.
2. We perform an in-depth investigation of randomness factors, and their interactions, that influence the training stability to allow for easier study of other properties and to improve replicability.
3. We investigate the factors affecting the transferability to allow for more efficient use of learning with limited labelled data in various domains.

*Supervised by: **Maria Bielikova** and **Ivan Srba**, Kempelen Institute of Intelligent Technologies

2 Direction of Research Work

Randomness factors influencing stability. We define stability as a property that indicates what influence the *small scale and random* changes in the input *data* and *parameters* have on the outputs of the model. These small scale changes (limited in their effect) are a result of inherent *randomness* of the training process. Therefore, they cannot be completely controlled, but rather must be taken into consideration and mitigated. As such randomness may significantly affect other properties, we focus on it first to make findings from later experiments more unbiased. At first, we plan to identify the factors in the training process, where the randomness is introduced (we denote them as *randomness factors*). They may be related to: 1) randomness in data (data splits, order of data or noise); or 2) model-specific parameters (random initialisation and minor changes to specific hyperparameters). Secondly, to measure the influence of these factors, we will observe the behaviour of models by systematic exploration of their effects (e.g., measuring the variance of results for multiple data splits). Besides improving the replicability and allowing us to draw unbiased conclusions from other experiments, knowing the source of instability also leads to more efficient use of the models in practice. We can identify parts of training process we should focus on, while also allowing us to disregard parts with no considerable impact on the variability of output.

Factors influencing transferability. After investigating stability, we will investigate transferability to new tasks and datasets in multilingual text domain, where the labels are commonly spread across languages and tasks. Transferability indicates how the outputs of the model change, when different, but related *dataset* and *task definition* is used as a source of additional knowledge. At first, we will identify factors influencing transferability. Preliminary candidates are the *task similarity* and *similarity of data distribution*, although they may still be too broad and other factors may exist. Afterwards, we will investigate the behaviour around these factors. In case of the similarities, this will require defining a consistent similarity measure between tasks and datasets, and then using it to investigate the behaviour when the similarity changes. We do not expect to identify one model that can deal with all the settings (although some will perform better), but this should help us determine what factors of transferability to focus on and how to control them to achieve better results, making them better suited for the specifics of our domain.

Interactions between factors. To account for the interactions, we will perform hierarchical investigation, changing values for each investigated parameter in the experiment, while keeping other values fixed. For evaluation, we will use more advanced statistical tests and models, such as Linear Mixed Model, in a similar fashion to [Boquet *et al.*, 2019].

Focus of experiments. We plan to focus on representative, most popular approaches for learning with limited labelled data that perform sufficiently well when combined with deep learning. First, we will focus on approaches utilizing related datasets and tasks, using meta-learning and transfer learning. Afterwards, our focus will shift to approaches that utilize unlabelled data, namely weakly supervised and semi-supervised learning.

3 Preliminary Results and Conclusion

We have performed a preliminary study of the stability for the optimisation based meta-learning approaches on a text sentiment analysis task. We have investigated the effects of two sets of factors - *data splits*, and the random *model initialisation*, *task sampling* and *sample order* in one - by running the training and evaluation on different train-validation-test splits, multiple times on each without fixing the random seeds. We used three basic meta-learning models: Model Agnostic Meta-Learning (MAML), its first order version (FO-MAML), and Reptile. To explore the behaviour when different number of labels is used, we also varied the amount of available labels. The results already show difference between approaches, with Reptile showing almost no variance contributed by the repeated runs, when running with low number of labels and optimized parameters, while being strongly affected by the choice of parameters. On the other hand, MAML shows consistent performance, only being outperformed by Reptile in a single case.

We have also applied transfer learning and the Reptile approach on a stance detection task in the domain of false information detection. Both approaches show promising results, significantly increasing performance on both datasets. However, the instability of results was noticeable, giving us further incentive for in-depth study of stability and transferability.

To conclude, in contrast to the existing approaches, we perform an in-depth investigation of the factors influencing stability and transferability to better control for their effects. Knowing and controlling these factors gives us good promise for success, as the improved reproducibility will allow us to draw unbiased conclusions from experiments.

Acknowledgements

This work was partially supported by The Ministry of Education, Science, Research and Sport of the Slovak Republic, Contract 0827/2021 and by the Central European Digital Media Observatory (CEDMO), Contract 2020-EU-IA-0267.

References

[Boquet *et al.*, 2019] Thomas Boquet, Laure Delisle, Denis Kochetkov, Nathan Schucher, Boris N Oreshkin, and Julien Cornebise. Reproducibility and stability analysis in metric-based few-shot learning. *RML@ ICLR*, 2019.

[Hospedales *et al.*, 2022] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2022.

[Lu *et al.*, 2021] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

[Ye *et al.*, 2021] Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021.