JYX JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Toivanen, Ida; Lindroos, Jari; Räsänen, Venla; Taipale, Sakari

Title: Dealing with a small amount of data : developing Finnish sentiment analysis

Year: 2022

Version: Accepted version (Final draft)

Copyright: © 2022 IEEE

Rights: In Copyright

Rights url: http://rightsstatements.org/page/InC/1.0/?language=en

Please cite the original version:

Toivanen, I., Lindroos, J., Räsänen, V., & Taipale, S. (2022). Dealing with a small amount of data : developing Finnish sentiment analysis. In 2022 BESC : 9th International Conference on Behavioural and Social Computing. IEEE. https://doi.org/10.1109/besc57393.2022.9995536

Dealing with a small amount of data – developing Finnish sentiment analysis

Ida Toivanen
Faculty of Humanities and Social Sciences
University of Jyvaskyla
Jyväskylä, Finland
ida.m.toivanen@jyu.fi

Jari Lindroos
Faculty of Humanities and Social Sciences
University of Jyvaskyla
Jyväskylä, Finland
jari.m.m.lindroos@jyu.fi

Venla Räsänen
Faculty of Humanities and Social Sciences
University of Jyvaskyla
Jyväskylä, Finland
venla.s.m.rasanen@jyu.fi

Sakari Taipale
Faculty of Humanities and Social Sciences
University of Jyvaskyla
Jyväskylä, Finland
sakari.taipale@jyu.fi

Abstract

Sentiment analysis has been more and more prominently visible among all natural language processing tasks. Sentiment analysis entails information extraction of opinions, emotions, and sentiments. In this paper, we aim to develop and test language models for low-resource language Finnish. We use the term "low-resource" to describe a language lacking in available resources for language modeling, especially annotated data. We investigate four models: the state-of-the-art FinBERT [1], and competitive alternative BERT models Finnish ConvBERT [2], Finnish Electra [3], and Finnish RoBERTa [4]. Having a comparative framework of multiple BERT variations is connected to our use of additional methods that are implemented to counteract the lack of annotated data. Basing our sentiment analysis on partly annotated survey data collected from eldercare workers, we supplement our training data with additional data sources. In addition to the non-annotated section of our survey data, additional data (external in-domain dataset and open-source news corpus) are focused on to determine how training data can be increased with the use of methods like pretraining (masked language modeling) and pseudo-labeling. Pretraining and pseudo-labeling, often defined as semi-supervised learning methods, make it possible to utilize unlabeled data either by initializing the model, or by labeling unlabeled data samples with seemingly real labels prior to actual model implementation. Our results suggest that out of all the single BERT models, FinBERT performs the best for our use case. Moreover, applying ensemble learning and combining multiple models further betters model performance and predictive power, and it outperforms a single FinBERT model. The use of both pseudo-labeling and ensemble learning proved to be valuable assets in the extension of training data for low-resource languages such as Finnish. However, with pseudo labeling, proper regularization methods should be considered to prevent confirmation bias from affecting the model performance.

Keywords: sentiment analysis, low-resource language, pseudo-labeling, BERT, ensemble learning

I. Introduction

Survey answers are a popular way of gathering information on a wide range of subjects. While pre-determined scales with fixed response choices, such as dichotomic, multiple choice, or Likert scales, might produce more maintainable data, open-ended questions tend to generate more versatile and rich information. Open-ended questions are particularly suitable when surveying novel topics that lack standardized and validated measuring scales. The problem with the extensive use of open-ended questionnaires is the amount of work required to manually analyze the gathered data. In recent years, multiple studies have shed light on the possibilities that sentiment analysis offers for this field. The benefit of using sentiment analysis in a straightforward way of dividing the answers into categories (e.g., negative, positive, and neutral) is that, in addition to an overview of the distribution of the sentiments, it gives us a possibility to reflect on specific answers for each category.

The aim of our study is to develop Finnish language models. We use survey answers from workers in the eldercare sector to implement and train these models. Questions on the relationship between traditional healthcare and the digitalization of healthcare practices bring out opinions that are often enriched with strong sentiments. Although assistive technologies are built to improve the healthcare sector's efficiency and safety, there is a chance that when digitalization is executed in a hurry and without consulting primary users, the results can be close to catastrophic [5][6]. The digitalization of healthcare services has already grown rapidly, and the COVID-19 pandemic has made the need for innovative and new technologies even more evident. These technologies might bring challenges to workers in the healthcare sector, since in addition to their preliminary work, they must adapt to a digitally changing work environment. Scholars have approached the subject in several ways [6][7][8][9], which suggests that sentiment analysis in this context should be properly researched. All of this makes the subject a rich source of data for our language model testing. The healthcare sector offers us an opportunity to test models with a field-specific vocabulary, and in this way, an opportunity to contribute not only to the development of Finnish textual models, but to the analysis of open-ended survey answers as well as to a field that prominently and continuously influences society.

We test different models and their ensembles to find the most competent ones for low-resource languages. We utilize pseudo-labeling and enhanced training data revolving around the topic of digitalization in healthcare services. The data used in our research predominantly involves the sentiments voiced out by eldercare workers who need to use digital services and technological devices in their work. Along with a University of Jyväskylä survey study on eldercare work [10] [11] collected in 2019 and 2021, we make use of scraped external data related to the same topic and the Digitoday 2014 corpus

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

[12] containing technology news data. With low-resource languages such as Finnish, there is a lack of data suitable for model training. For this reason, we decided to grow our training data by collecting external data and searching for open data that are contextually close to our topic (i.e., technology news data) for inclusion in the model implementation process.

Finnish language modeling is moving forward despite the data challenges; simultaneously, the usable resources within this field continue to grow. Annotated data being scarce, we intend to use alternative measures for additional data to expand the training data for language models and improve their performance. These measures include pretraining (by using masked language modeling) and pseudo-labeling, which refers to the forming of artificial labels for unlabeled data samples. After utilizing these methods to increase training data, we redirect our focus on BERT-based modeling and to different open-source backbones: FinBERT [1], Finnish ConvBERT [2], Electra [3], and RoBERTa [4]. Training these single models and ensembling them to include predictions from multiple models offers us a wide research base for the chosen task of sentiment analysis. Thus, this paper answers the following two research questions:

I. Do pseudo-labeling and ensemble learning improve model performance when working with little annotated data?

II. Which model performs best with the given sentiment analysis task?

II. Literature Review

Sentiment analysis is a natural language processing (NLP) task that involves the studying of public opinion, usually to attain information about the subjective attitudes of people and obtain user feedback on systems and services [13]. These public opinions can be classified as positive, negative, or neutral depending on the type of sentiments they convey [13]. Currently, state-of-the-art models on sentiment analysis benchmarks, such as Internet Movie Database (IMDb) [31], are dependent on language modeling methods that are based on large amounts of data, making it hard to advance modeling for languages that lack those resources.

Low-resource language is a term used to refer to the type of language that is lacking in available data, especially data that are annotated and/or in a digital form [14][15]. In NLP, in addition to more prevalent supervised learning that deals with annotated (labeled) data, there has been a surge of methods for low-resourced languages that has allowed the extension of language modeling to a wider range of languages. These methods may be described as semi-supervised, which can roughly be defined as methods that use both unlabeled and labeled data samples [16]. One of these semi-supervised methods is pseudo-labeling [17][18].

Pseudo-labeling refers to the forming of labels to new and unlabeled data by obtaining predictions from a model that has been trained with labeled data. More precisely, let us follow the notation presented in [19], in which a model $f_\theta(x)$ is trained with a training dataset $\mathcal{T}$ of $M = M_u + M_l$ samples that consist of an unlabeled dataset $\mathcal{T}_u = \{x_i\}_{i=1}^{M_u}$ and a labeled dataset $\mathcal{T}_\ell = \{(x_i, y_i)\}_{i=1}^{M_l}$. The one-hot encoding labels $y_i \in \{0,1\}^C$ for C classes correspond to x_i; when applying pseudo-labeling for M_u unlabeled data samples there are pseudo labels $\tilde{y}$ available. The training dataset, now including labels for all the data, can then be refined as $\tilde{\mathcal{T}} = \{(x_i, \tilde{y}_i)\}_{i=1}^{M}$ where $\tilde{y} = y$ for the labeled data samples M_l.

Besides pseudo labeling and the harnessing of unlabeled data via means that imitate supervised learning, unlabeled data can be taken into use in an unsupervised way by pretraining. Pretraining is a valuable option for leveraging information in low-resource languages that tend to lack labeled but not unlabeled data [20]. Transformer-based BERT modeling has paved its way into deep learning research, introducing masked language modeling as a pretraining method [21]. In masked language modeling (MLM), a certain portion of the input tokens are masked randomly to attain information bidirectionally [21]. The objective of MLM is to have the mask [MASK] replace the token z_t. This is then predicted based on the information on past and future tokens, which can be defined as $Z_{\setminus t} := (z_1, \dots, z_{t-1}, z_{t+1}, \dots, z_{|Z|})$ [22].

Pseudo labeling and BERT modeling have been successfully combined in language modeling research. For example, offensive language in low-resourced Dravidian languages (Tamil, Kannada, and Malayalam) was identified with the help of pseudo labeling [23]. Pseudo labeling in this case was used to increase the amount of training data. For the NLP task, pretrained BERT models (multilingual BERT, IndicBERT, DistilBERT) along with a transformer-based XLM model were tested against ULMFIT [24]. In another study [25] using data from StackOverflow, different variations of DistilBERT were compared in a question-answering task. The study concluded that DistilBERT, which utilized pseudo-labeled data, outperformed other model variations implemented in the study.

Instead of the use of BERT models in conjunction with pseudo labeling, the effect of adding transfer learning and pseudo labeling separately to the model implementation process has been studied. The benefits of pseudo labels and transfer learning were studied previously, and it was shown that models using pseudo labeled data brought competitive results when compared to transfer learning models—especially when additional finetuning was applied after training with pseudo labels [26]. In this case, the study focused on neural ranking tasks. Pseudo labeling seems to have potential in terms of adding new data for training, but the tendency of pseudo labeling to increase confirmation bias in the model implementation can be disadvantageous [19]. To counteract this, regularization methods, such as dropout, can be used in the model training process.

To reduce generalization errors, ensemble methods may be considered. Ensemble methods involve the use of multiple models that are combined to better the model performance since models tend not to give the same errors on the same test dataset—thus, the models complement each other [16]. Ensemble methods typically involve training several models on the same dataset and then combining each prediction of the models into one final prediction. Ensemble methods have increasingly been implemented in machine learning contests, where participants compete in model implementation. This has led to the winning parties mostly implementing a combination of a larger set of models rather than a single model [16]. Recently, in addition to machine learning competitions (like those presented on the Kaggle platform), the effectiveness of ensembling BERT-based models has been researched [27][28]. The utilization of ensemble learning has

proven to be effective in the detection of offensive language in English, Hindi, and Marathi text data [28] and in news identification for text sentiment classification [27].

III. Experimental Data

We use the following data in the model implementation: 1) survey studies on eldercare work, 2) external data related to our domain, and 3) online technology news data. The eldercare survey data [10][11] consists of answers to two different open-ended questions: 1) "What kind of emotions related to the use of technology have been present in your work during the last week?" ("Millaisia tunteita teknologian käyttämiseen on liittynyt työssä viimeisen viikon aikana?"), and 2) "What do you think about the following claim: 'Technology improves the quality of eldercare work and decreases the pressure of employees'?" ("Mitä ajattelet väitteestä: 'Teknologia parantaa vanhustyön laatua ja vähentää työntekijöiden kuormitusta'"). This survey data (1st dataset) was collected in 2019 and 2021 by the Centre of Excellence in Research on Ageing at the University of Jyväskylä. The aim of the survey was to collect information on the working conditions and use of information communication technologies (ICTs) among eldercare workers in Finland. In addition to eldercare survey data, we apply external in-domain data (2nd dataset), which was scraped with queries that have topics like the eldercare survey data (such as "eldercare", "technology", and "digitalization"). Finally, we use the Digitoday 2014 corpus [12] (3rd dataset) that consists of Finnish technology news. The eldercare survey data amounts to 8274 respondents, out of which the answers for the first open-ended question (n = 4030) are annotated to three classes (negative, positive, and neutral), and answers for the second open-ended question (n = 4244) have no annotations. External in-domain data amounts to 289 data samples, and the Digitoday news corpus amounts to 14,483 data samples. Both datasets, the external and Digitoday news corpus, do not include annotations for the classes negative, positive, and neutral.

The data sample sizes (i.e., the number of tokens) differ with each dataset that we use. In eldercare survey data, the data sample sizes for the answers for the first open-ended question range from 3 (minimum) to 429 (maximum) tokens, with a median length of 13 and a mean of 18.9 tokens. The data sample sizes for the answers for the second open-ended question range from 3 (minimum) to 658 (maximum) tokens, the median length being 16 and the mean 23.3 tokens. The external dataset has samples in the size range of 10 (minimum) to 116 (maximum) tokens, while the median length is 36 and the mean length is 40.6 tokens. After preprocessing (see Chapter IV, Section A), we find that the Digitoday corpus has data samples from the size of 3 (minimum) up to 658 (maximum) tokens. For the same corpus, a median length of 19 and a mean length of 21.6 tokens are obtained.

IV. Methodology

A. Preprocessing

During data preprocessing, empty and duplicate values were removed from the data, and Swedish answers were removed from the eldercare survey data. The format of the Digitoday corpus was changed to better suit our purposes: excluding annotations for named entity recognition and separating data samples by sentences rather than words. After these preprocessing steps, the test dataset (n = 806) was extracted from the eldercare survey data (first open-ended question) for the testing of all model variations.

TABLE I. Class distribution for the training data before and after pseudo labeling.

Training data	*Positive*	*Negative*	*Neutral*
Before adding PL data	12.78%	55.46%	31.76%
After adding PL data	12.01%	34.10%	53.89%

B. Baseline and pseudo labeling

We decided to build the backbone of the baseline on the state-of-the-art Finnish BERT model FinBERT [1]. After implementation of the baseline, we used it for pseudo labeling. We applied pretraining with a masked language modeling (MLM) objective with all the data for the baseline. We then fine-tuned the pretrained model with the annotated eldercare data. This baseline was then used to give pseudo labels for the rest of the eldercare survey data (answers to the second open-ended question) and the external in-domain data. The Digitoday corpus connects to our subject in the field of technology, but since our main topic (digitalization in eldercare) was reasonably different, the pseudo labeling of the Digitoday corpus became unnecessary. Pseudo labels were given to 4533 data samples, of which 520 were positive, 857 negative, and 3156 neutral. For reference, out of the 3224 training data samples in the annotated eldercare data, 412 were positive, 1788 were negative, and 1024 were neutral. After adding pseudo-labeled data into the training data, instead of negative samples being more prevalent, the neutral data samples were more prominent in the final training dataset (see Table I). The baseline training and the pseudo labeling process connecting to the implementation of the rest of the language models is visualized in Fig. 1.

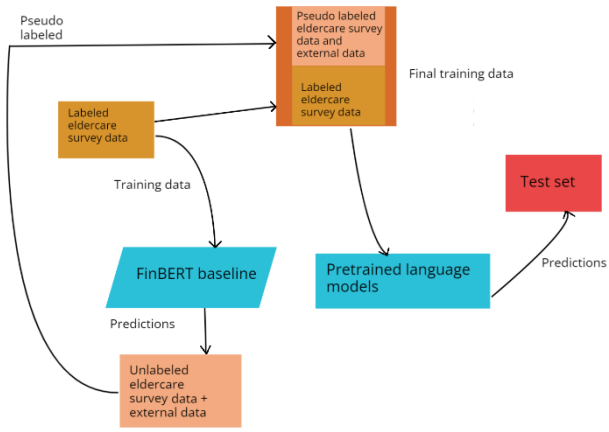

Fig.1. The construction of baseline for pseudo labeling and the use of different datasets.

C. Model implementation

We built model variations on four different language model backbones for model implementation. These backbones were FinBERT (110M parameters) [1], Finnish ConvBERT (106M parameters) [2], Finnish Electra (110M parameters) [3], and Finnish RoBERTa (355M parameters)

[4]. FinBERT was chosen because it is the current state-of-the-art Finnish language model for many NLP tasks [1]. Every Finnish model variation has been pretrained on large amounts of raw Finnish text in a self-supervised fashion, with the masked language modeling objective for FinBERT and RoBERTa and replaced token detection objective for Electra and ConvBERT.

To test FinBERT's usability in the sentiment analysis task, we also chose three architecturally newer models for comparison. The newer models, Finnish ConvBERT, Finnish Electra, and Finnish RoBERTa, offered a diverse base to achieve this objective. All the models underwent the same implementation process:

1. Pretraining with masked language modeling and with all our data,
2. using annotated and pseudo-labeled data (for all datasets except Digitoday) to train a finetuned model that was initialized with the pretrained model weights, and
3. applying ensemble learning.

All these steps were done with the same fixed seed, and with each model, the corresponding tokenizer was initialized (with the FinBERT model, the FinBERT tokenizer, etc.). In the pretraining phase for MLM, we used the following hyperparameters: maximum sequence length (number of tokens) of 110, batch size of 64 for training and validation (except for the RoBERTa model, which had a batch size of 32), AdamW optimizer with initial learning rate of 1e-6, linear scheduler with no warmup steps and a weight decay of 0.01, and masked language modeling with a probability of 0.15. We trained each pretrained model for three epochs. We conducted finetuning with hyperparameters of maximum sequence length of 100, batch size of 32 for training and 64 for validation, an AdamW optimizer with an initial learning rate of 5e-4, and a cosine scheduler with no warmup steps. Layer-wise learning rate decay [29] was used for the layers of every model; this ranged from 3e-5 to 1e-5. We trained each model for 8 epochs and used stratified K-fold cross validation (K = 5) to split the data into five groups.

On top of each model's backbone, we included additional layers. The architecture consisted of a linear layer, a pooling layer, for calculating the weighted mean of the different hidden layer representations of token embeddings, multi-sample dropout layers, and the classification layer for the corresponding three sentiment classes.

To determine whether validation loss or validation F1-score is a more suitable validation metric for choosing the out-of-fold models and for estimating the performance of the models on the test set, we incorporated the best folds of both fold metrics to be further evaluated on the test set for every model variation. The three fold metrics chosen were validation loss, validation F1, and the combined predictions by both validation loss and validation F1. This means that for validation loss, five folds were averaged out; for validation F1-score five folds were averaged out. For the combination of validation loss and F1-score, a total of 10 folds were averaged out. Finally, the models were ensembled for each fold metric by averaging the single predictions from each model's output layer to output one final prediction for the labels of the test set. Incorporating multiple different backbones ensures diverse predictions for lower bias and variance.

V. Results

The results of the sentiment classification task were obtained with several evaluation metrics: accuracy, weighted F1-score, and precision. All results obtained with the evaluation metrics are shown in Table II for the baseline model, which was used for pseudo labeling the unlabeled eldercare survey data and external data, Table III for the single models, and Table IV for the corresponding ensembled models. The baseline model, shown in Table II, obtained an F1-score of 89.21%, which was used for pseudo labeling the new training data for every following model in the results presented in Table. III and Table. IV.

The results suggest that FinBERT produces the best results of all the single models with the combined fold metric for every evaluation metric, with an F1-score of 90.46%. The increase in the F1-score over the baseline was 1.02%. The second-best single model, ConvBERT, that used a F1-score fold metric, also showed a 0.99% increase in the F1-score compared to the baseline. The other models, Electra and RoBERTa, seemed to perform worse than the FinBERT and ConvBERT models. The difference in the F1-score for the Electra and RoBERTa models when compared to the FinBERT model was already around 1%. There was no single distinguishable fold metric that could be chosen overall for different single models. For example, when comparing the results of different single models in terms of the same fold metric, based on the F1-score, the results do not change in a relative sense when using varied fold metrics. In other words, FinBERT still outperforms (with F1-score) other models whether the fold metric is loss, F1-score, or the combination of loss and F1-score. Looking at our results and the difference between the results with different fold metrics of one model, we suggest that it is safer to choose the out-of-fold models by considering more than a single criterion.

The combination of FinBERT and ConvBERT for the ensembled models with the validation loss as the fold metric produced the overall best score for the sentiment classification, with an F1-score of 90.97%. This could be due to the complementary nature of FinBERT and ConvBERT models; in addition, they are the best-performing single models. This represents intriguing possibilities, because the ensembled models' single counterparts are not the best performing single models when validation loss is the fold metric; however, their predictions suggest that they are the best here when they are combined. Ensembling shows an increase in performance for every listed variation, as shown in Table IV, compared to the single models, except for the ensemble of FinBERT, ConvBERT, Electra, and RoBERTa models with the validation loss fold metric. No clear single-fold metric was the best for ensembling, but in two out of three cases, the best fold metric was validation loss.

TABLE II. Results for the baseline model for the test set.

Model	***Accuracy***	***F1-score***	***Precision***
FinBERT	89.21	89.21	89.25

TABLE III. Results for single models with pseudo labeling on the test set.

Model	**Fold metric**	***Accuracy***	***F1-score***	***Precision***

FinBERT	Loss	90.20	90.23	90.36
	F1	90.20	90.20	90.24
	Loss+F1	**90.45**	**90.46**	**90.54**
ConvBERT	Loss	89.83	89.80	89.83
	F1	**90.20**	**90.16**	**90.17**
	Loss+F1	90.07	90.04	90.06
Electra-base	Loss	89.21	89.16	89.20
	F1	88.34	88.29	88.32
	Loss+F1	**89.45**	**89.43**	**89.44**
RoBERTa-large	Loss	**88.59**	**88.55**	**88.69**
	F1	88.33	88.30	88.32
	Loss+F1	88.46	88.43	88.48

TABLE IV. RESULTS FOR THE ENSEMBLED MODELS WITH PSEUDO LABELING ON THE TEST SET.

Model	**Fold metric**	***Accuracy***	***F1-score***	***Precision***
FinBERT + ConvBERT-base	Loss	**90.94**	**90.97**	**91.10**
	F1	90.70	90.70	90.77
	Loss+F1	90.70	90.70	90.74
FinBERT + ConvBert-base + Electra-base	Loss	90.82	**90.83**	**90.88**
	F1	90.82	90.82	90.87
	Loss+F1	90.82	90.81	90.84
FinBERT + ConvBert-base + Electra-base + RoBERTa-large	Loss	90.32	90.32	90.36
	F1	90.82	90.80	90.83
	Loss+F1	90.82	**90.82**	**90.85**

VI. DISCUSSION

Low-resource languages are the type of languages that, when in use, would extensively benefit merely from using more data. Although we had manually annotated in-domain data on a specific topic at hand—technology use in eldercare work—for which sentiment analysis can directly be implemented, the rest of our data was not annotated and was related on a more implicit level to our topic. The topic is still, by and large, at the core of the area of eldercare data that consists of the answers for the second open-ended question, but it is decidedly different in the sense that the second open-ended question is not purely asking about the sentiments voiced out by the survey responders, but instead making a statement to be commented about. Most of the answers to this question naturally only concisely agreed or disagreed with the presented statement (e.g., with "Yes"), sometimes adding a justification. When comparing the number of positive, negative, and neutral data samples (see Table I) we can notice that after adding the pseudo-labeled data into the training data, there is a shift from negative to neutral class that accounts for the largest amount of data samples. Observing changes such as this in data distribution, it is important to note that, when using survey data, the way the survey questions are presented directly affects the answers obtained. Utilizing data scrapes on a larger scope—such as, for example, on several social media platforms—would enable us to obtain more public opinion data with variability between positive and negative classes.

Pseudo labeling is a semi-supervised method that is a popular way to add up more data without needing to manually annotate data. In our case as well, pseudo labeling the data had a positive effect on the outcome of the models and precipitated a great increase in performance for FinBERT (see Table II or the baseline trained without pseudo-labeled data; see Table III for FinBERT that was trained including pseudo-labeled data). Although pseudo labeling may prove to be useful, possible confirmation bias (due to noisy labels) suggests that certain measures should be conducted when implementing pseudo labeling. For example, with regularization methods, confirmation bias can be reduced [19]. In our case, the use of multi-sample dropout, a type of regularization method in which units in the neural network layers are randomly dropped, and pooling layer for calculating the weighted mean of the different hidden layer representations helped regularize the model. These architectural choices help reduce the possible negative effect noisy labels could have on our models. Additionally, the use of iterative pseudo labeling, in which we pseudo label the data more than once with the continuously trained models, could be something to investigate on, as it has been shown to give superior results in both standard and low-resource settings, compared to the conventional pseudo labeling approaches [30].

Another means of producing better predictions lies in the use of ensemble learning. Ensemble learning is a technique for combining multiple predictions of different models trained on the same data. With only one exception, our ensembled models showed an increase in performance compared to the single models. It should be noted that our ensembles consist of a modest number of models—the ensembling of models can be extended so that predictions are produced using tens of different models, for instance. Additionally, we demonstrated that choosing out-of-fold models using different criteria can have an impact on the results. In most cases, selecting the best performing model based on a certain criterion does not always mean it is the best performing one—in this case, the combined fold metrics or the validation loss metric had the best performance among the models. Regardless of the several methods that could have been further improved on in this piece of research, the automatic identification of sentiment content remains an important application for the Finnish language among other low-resource languages.

VII. CONCLUSIONS

This study was carried out to obtain more information about the methods that can be used to increase training data in low-resource languages. In this case, we used Finnish as a representative case of a language that lacks resources, such as openly available data and annotated data for different NLP tasks. Using language modeling to understand the Finnish language, we focused on BERT-based modeling in the NLP task of sentiment analysis. The sentiment analysis was mainly

conducted using survey data that contain sentiments regarding the use of technology and digitalization made by eldercare workers. In addition to survey data, and to overcome the limitation of not having enough data on hand, we enhanced the training data with an externally scraped dataset and with an open dataset involving technology news in Finnish. Since the enhanced training data was not fully annotated, we attempted to utilize pseudo labeling and ensemble learning to enable the use of datasets that were not annotated, and to see what the effects would be using a model averaging method like ensemble learning in our case. Considering all of this, we aimed to answer two research questions: I) "Do pseudo labeling and ensemble learning improve the model performance when working with little annotated data?", and II) "Which model performs the best with the given sentiment analysis task?".

To approach these research questions, we first trained a baseline model with annotated data to utilize pseudo labeling for non-annotated data. Using the backbones FinBERT, Finnish ConvBERT, Finnish Electra, and Finnish RoBERTa, we trained several single models utilizing the pseudo-labeled data and annotated data in order to draw comparisons between models. To answer the first research question, we found that labeling seems to improve the model performance of FinBERT (see Table III) when compared to the baseline FinBERT (see Table II) for which pseudo-labeled data was not used. In addition to training multiple single models, we ensembled models to contain and link predictions from multiple models to determine whether the predictions would prove to be more accurate. To answer the first research question, our results suggest that ensembling further improves the model's predictive power when compared to single models' performance (see Tables III and IV). To answer the second research question, we compared the results of models with different backbones. The results shown in Table III suggest that FinBERT produces the best results out of all the single models we implemented. When single and ensembled models are all compared to each other, the ensemble of FinBERT and ConvBERT is, overall, the best model obtained in our study.

Due to the lack of trained tools for low-resource languages, there remains a need to deploy more comprehensive sentiment analysis tasks in the field of humanities and social sciences for more than just detecting sentiments about the use of technology in the eldercare context. This should be taken into account when designing and implementing models for sentiment analysis in future work.

REFERENCES

[1] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, ... and S. Pyysalo, "Multilingual is not enough: BERT for Finnish," 2019. [Online]. Available: https://arxiv.org/abs/1912.07076

[2] A. Tanskanen and R. Toivanen. (2022b). ConvBERT for Finnish. Huggingface model. From https://huggingface.co/Finnish-NLP/convbert-base-finnish

[3] A. Tanskanen and R. Toivanen. (2022a). ELECTRA for Finnish. Huggingface model. From https://huggingface.co/Finnish-NLP/electra-base-discriminator-finnish

[4] A. Tanskanen, R. Toivanen and T. Vehviläinen. (2022). RoBERTa large model for Finnish. Huggingface model. From https://huggingface.co/Finnish-NLP/roberta-large-finnish

[5] European Digital Agenda, "Digital Agenda for Europe," 2022. Accessed March 21, 2022. [Online]. Available: https://www.europarl.europa.eu/ftu/pdf/en/FTU2.4.3.pdf

[6] A. Bartosiewicz, J. Burzyńska and P. Januszewicz, "Polish Nurses' Attitude to e-Health Solutions and Self-Assessment of Their IT Competence," *Journal of Clinical Medicine*, vol. 10, no. 20 p. 4799, 2021.

[7] M. Laitinen, T. Hantunen, T. Heino, P. Hilama, A. Huttunen, P. Janhunen, . . . K. Ammattikorkeakoulu, "Digi vie, sote vikisee": Kokemuksia sote-alan digitalisaatiosta DigiSote-hankkeessa Etelä-Savossa ("The digital takes, the health and social services complain": Experiences about the digitalization of health and social services in DigiSote project in South Savo."), 2018. Accessed July 17, 2022. [Online]. Available: https://urn.fi/URN:ISBN:978-952-344-090-6

[8] K. Ylönen, S. Salovaara, J. Kaipio, M. Tyllinen, E. Tynkkynen, S. Hautala and T. Lääveri, "Sosiaalialan asiakastietojärjestelmissä paljon parannettavaa: käyttäjäkokemukset 2019 ("There is much to be improved on the client information systems in social services: user experiences in 2019")," *Finnish Journal of eHealth and eWelfare*, vol. 12, no. 1, pp. 30–43, (2020).

[9] K. Seibert, D. Domhoff, K. Huter, K., Krick, T., Rothgang, H. and K. Wolf-Ostermann, "Application of digital technologies in nursing practice: results of a mixed methods study on nurses' experiences, needs and perspectives," Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen, vol. 158, pp. 94–106, 2020.

[10] J. Karhinen, S. Taipale, M. Tammelin, A. Hämäläinen, H. Hirvonen and T. Oinas, "Vanhustyö ja teknologia. Jyväskylän yliopiston vanhustyön kyselytutkimus 2019: Katsaus tutkimusaineistoon ("Eldercare and technology. The survey study on eldercare by the University of Jyväskylä in 2019: A look on to the research data")", 2019. [Online]. Available: https://jyx.jyu.fi/handle/123456789/65649

[11] J. Karhinen, T. Oinas, M. Tammelin, A. Hämäläinen, H. Hirvonen, V. Mustola, E. Rantala and S. Taipale, "Vanhustyö ja teknologia. Jyväskylän yliopiston vanhustyön kyselytutkimus 2021: Katsaus tutkimusaineistoon ("Eldercare and technology. The survey study on eldercare by the University of Jyväskylä in 2021: A look on to the research data")", 2021. [Online]. Available: https://jyx.jyu.fi/handle/123456789/78742

[12] T. Ruokolainen, P. Kauppinen, M. Silfverberg, M. and K. Lindén, "A Finnish News Corpus for Named Entity Recognition," *Language Resources and Evaluation*, vol. 54, no. 1, pp. 247–272, 2020. Available: https://doi.org/10.1007/s10579-019-09471-7

[13] N. C. Dang, M. N. Moreno-García and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, p. 483, 2021. Available: https://arxiv.org/ftp/arxiv/papers/2006/2006.03541.pdf

[14] C. Cieri, M. Maxwell, S. Strassel and J. Tracey, "Selection criteria for low resource language programs," In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), pp. 4543–4549, 2016. Available: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/lrec2016-selection-criteria-for-low-resource-lang.pdf

[15] A. Magueresse, V. Carles and E. Heetderks, "Low-resource languages: A review of past work and future challenges," 2020. [Online]. Available: https://arxiv.org/pdf/2006.07264.pdf

[16] I. Goodfellow, Y. Bengio and A. Courville (2016). *Deep Learning*. MIT Press. Available: https://www.deeplearningbook.org/

[17] X. Yang, Z. Song, I. King and Z. Xu, "A survey on deep semi-supervised learning," 2021. [Online]. Available: https://arxiv.org/pdf/2103.00550.pdf

[18] J. E. Van Engelen and H. H, Hoos, "A survey on semi-supervised learning,". *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020. Available: https://link.springer.com/article/10.1007/s10994-019-05855-6

[19] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, N. E. and K. McGuinness, "Pseudo labeling and confirmation bias in deep semi-supervised learning," In *2020 International Joint Conference on Neural Networks* (IJCNN), pp. 1–8. IEEE. Available: https://ieeexplore.ieee.org/document/9207304

[20] J. C. B. Cruz and C. Cheng, "Establishing baselines for text classification in low-resource languages," 2022. [Online]. Available: https://arxiv.org/pdf/2005.02068.pdf

[21] J. Devlin, M.W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," In

Proceedings of NAACL-HLT 2019, pp. 4171–4186. 2019. Available: https://aclanthology.org/N19-1423.pdf

[22] J. Salazar, D. Liang, T. Q. Nguyen and K. Kirchhoff, "Masked language model scoring," In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2699–2712. 2020. Available: https://aclanthology.org/2020.acl-main.240.pdf

[23] A. Hande, K. Puranik, K. Yasaswini, R. Priyadharshini, S. Thavareesan, A. Sampath, A., ... and B.R. Chakravarthi, "Offensive language identification in low-resourced code-mixed dravidian languages using pseudo labeling," 2021. [Online]. Available: https://arxiv.org/pdf/2108.12177.pdf

[24] J. Howard, J. and S. Ruder, "Universal language model fine-tuning for text classification," In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 328–339. 2018. Available: https://aclanthology.org/P18-1031.pdf

[25] K. Kuligowska and Kowalczuk, B. (2021). "Pseudo labeling with transformers for improving Question Answering systems," *Procedia Computer Science*, vol. 192, pp. 1162-1169. Available: https://www.sciencedirect.com/science/article/pii/S1877050921016082

[26] I. Mokrii, L. Boytsov and P. Braslavski, "A systematic evaluation of transfer learning and pseudo labeling with bert-based ranking models," In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2081–2085. 2021. Available: https://dl.acm.org/doi/pdf/10.1145/3404835.3463093

[27] S. Y. Lin, Y. C. Kung and F.Y. Leu, "Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis," *Information Processing and Management*, vol. 59, no. 2, 102872, 2022. Available: https://www.sciencedirect.com/science/article/abs/pii/S0306457322000073

[28] C. Caparrós-Laiz, J, Antonio, G. Díaz and R. Valencia-Garcia, "Detecting Hate Speech on English and Indo-Aryan Languages with BERT and Ensemble Learning," In *Forum for Information Retrieval Evaluation* (FIRE) (Working Notes), CEUR-WS.org, 2021. Available: http://ceur-ws.org/Vol-3159/T1-7.pdf

[29] T. Zhang, F. Wu, A. Katiyar, K.Q. Weinberger and Y. Artzi, "Revisiting few-sample BERT fine-tuning," 2020. [Online]. Available: https://arxiv.org/abs/2006.05987

[30] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve and R. Collobert, "Iterative pseudo labeling for speech recognition," 2020. [Online]. Available: https://arxiv.org/pdf/2005.09267.pdf

[31] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng and C. Potts, "Learning word vectors for sentiment analysis," In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, pp. 142-150, 2018.