Murre24: Dialect Identification of Finnish Internet Forum Messages

Olli Kuparinen
Faculty of Information Technology and Communication Sciences, Tampere University
Department of Digital Humanities, University of Helsinki
first.last@tuni.fi

Abstract

This paper presents Murre24, a collection of dialectal messages posted on the largest Finnish internet forum, Suomi24. The messages posted in Finnish on the forum between 2001 and 2020 are classified to present either the standard language, one of the seven traditional dialects, a colloquial style or the Helsinki slang. We present a manually annotated dataset used to train dialect identification models as well as the automatic annotation of almost 94 million messages in total. We experiment with five different dialect identification methods and evaluate them on dialectally balanced and random test samples. The best performing method for differentiating standard Finnish from non-standard Finnish is a character n-gram based support vector machine (SVM), while fine-tuning a BERT-based model achieves best scores in the final dialect identification task. According to the automatic classification, most of the messages written on the forum are in standard Finnish, and most of the non-standard messages are in a colloquial variety used typically by young speakers in Finland. We moreover show that the proportion of non-standard messages declines over time, but the proportion of the traditional dialects stays relatively steady.

Keywords: dialect identification, user-generated content, Finnish

Spoken dialect corpora are typically very costly to produce, requiring expert knowledge and working hours to interview speakers and transcribe speech consistently. Dialectologists and sociolinguists have thus turned increasingly towards variation encountered in user-generated content (UGC) in social media to overcome this data bottleneck. Internet texts often contain non-standard language of several origins, such as dialectal forms, abbreviations and misspellings. A key problem, therefore, is how to extract the non-standard content from the large mass of texts. Language and dialect identification tools are often used for this purpose.

While several off-the-shelf tools exist for language identification (e.g., Lui and Baldwin, 2012; Joulin et al., 2016; Jauhiainen et al., 2022), dialect identification is typically a harder problem. This can be explained by the amount of available training data, and by the similarity of the language forms to be distinguished. The issue of dialect identification has been mostly examined in the shared tasks organized in the VarDial workshop, which have focused, for instance, on discriminating between Uralic languages (Jauhiainen et al., 2020), Swiss German dialects (Zampieri et al., 2017), and Italian dialects (Aepli et al., 2022).

This paper presents Murre24 ('Dialect24'), a collection of dialectal messages written in the largest Finnish internet forum, Suomi24 ('Finland24'). The messages posted on the forum between 2001 and 2020 have been published as a corpus with an academic license (City Digital Group, 2021a), and they total to almost 94 million messages. In this work, we identify the dialects used in the messages in a three-step process.[1]

Moreover, we create a dataset of manually annotated dialectal messages to be used in further work of dialect identification. The paper also discusses different identification methodologies and includes statistics of the variation and longitudinal change of dialect use in the forum.

The contributions of the paper are:

- a manually annotated dataset of around 4000 Finnish internet forum messages,
- an automatic annotation of all 94M messages in the forum corpus,
- an evaluation of five different language identification tools on the task, and
- a discussion on the variation and language change in the dataset.

1. Related Work

1.1. Collection of Dialectal User-generated Content

Building spoken language corpora is very time-consuming since both interviewing and especially transcribing are difficult tasks. As a result, strides have been made to collect dialectal user-generated content from the Web.

Many collection efforts have focused on Twitter, which has allowed researchers to collect tweets

[1]The dialectal annotations and code are published at https://github.com/Helsinki-NLP/murre24.

LREC-COLING 2024, pages 12003–12015

through their API. Ljubešić et al. (2016) collect tweets with a custom tool from the South Slavic language continuum (Bosnian, Croatian, Montenegrin, and Serbian), while Huang et al. (2016) study dialectal variation in tweets originating in the US. Mubarak (2018) describes the collection of dialectal Arabic tweets and their standardization to modern standard Arabic. Another Arabic dialect tweet compilation is described in Althobaiti (2022). Barnes et al. (2021) annotate Norwegian tweets by language variety, with dialect being one of them, and Kuparinen (2023) introduces a collection of dialectal Finnish tweets written during a "dialect week" on Twitter.

Collection from other social media platforms has not been as popular, but Ueberwasser and Stark (2017) collect WhatsApp messages from Switzerland, while Hovy and Purschke (2018) describe a collection of over 16 million Jodel posts from German-speaking areas. The MultiLexNorm (van der Goot et al., 2021) collection includes data from 12 languages with variation typical of social media: abbreviations, typos, and dialectal features. The data is mostly from Twitter, but other outlets are also used.

Especially Arabic dialectal content has been collected from online commentaries. Zaidan and Callison-Burch (2011) collect and annotate dialectal comments from three Arabic news outlets, while Salama et al. (2014) collect Arabic comments from Youtube.

1.2. Language and Dialect Identification

Automatic language identification is a process of attaching a language label to a text. The problem has been studied for decades, and has been declared solved for distant languages in long texts (McNamee, 2005). When the texts are short or the languages are similar, the task gets increasingly difficult. A comprehensive look on the subject is presented in Jauhiainen et al. (2019b).

Most text classification methods can be trained to identify languages. The most popular methods for language identification have been linear models, such as support vector machines (SVM), Naïve Bayes (NB) and logistic regression (Wu et al., 2019; Jauhiainen et al., 2019a; Camposampiero et al., 2022).

While neural models have risen to state-of-the-art in many NLP tasks, they have not fared as well in dialect identification. Convolutional Neural Networks (CNN; Zhang et al. 2015) were the go-to neural solution in the past (Ali, 2018), but Transformer-based neural models have since become more utilized. Bernier-Colborne et al. (2019) describe the building of a BERT-like model (Devlin et al., 2019), while Zaharia et al. (2020) fine-tune a pre-existing Romanian BERT model.

Finally, the off-the-shelf tools fastText (Joulin et al., 2016) and HeLI (Jauhiainen et al., 2022) can identify several languages on their own, but they can also be re-trained with custom data.

2. Data

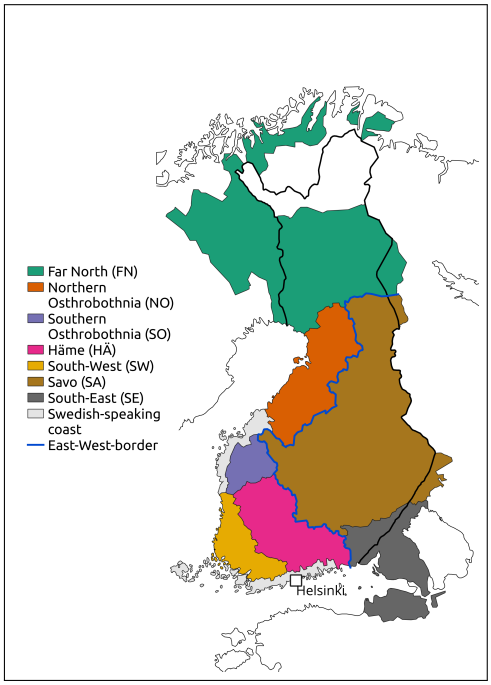

Figure 1: The seven dialect areas of Finnish and the capital Helsinki. The dialect areas are based on Itkonen (1989). Dialects are nowadays spoken mostly inside Finland (borders in black). The Northern Ostrobothnia in the map also includes Central Osthrobothnia. The Northernmost areas of Finland are the Sámi homeland.

2.1. Suomi24

Suomi24 is the largest internet forum in Finland and it is owned by City Digital Group. The messages from 2001 to 2020 have been published in the Language Bank of Finland, where they are usable with an academic license (City Digital Group, 2021a). The corpus includes almost 94 million messages, most of which are written in Finnish (see Section 3.1 for details).

In the current work, we automatically identify the variety used in each message in the corpus. The resulting dialect corpus can be used for dialectological and sociolinguistic studies, or for building NLP models on non-standard data.

The messages are annotated in a three-step process: first identifying the messages in Finnish, then identifying the messages in non-standard Finnish, and finally identifying the dialects.

2.1.1. Manual Annotation

We manually annotate the language variety used in a subset of messages in the corpus. We then use the annotated messages as training data for an automatic dialect identifier, but they could also be used on their own in further work. This manually annotated dataset is hereafter referenced as **S24** to differentiate from the forum name.

We first searched for words containing dialectal features and collected the messages which were clearly written in a dialect. After collecting a set of messages for each dialect, we trained an SVM-based classification model on this data and predicted the dialects in a new set of messages. The correctly identified messages were kept from this stage to build the final manually annotated dataset.

Our annotation is based on a seven-way division of Finnish dialects, presented in Itkonen (1989). In the original work, an eighth dialect area, called transitional Southwestern dialects, exists between South-West and Häme. However, it shares many features with both dialects, and it would be hard to discern it from these in short messages. It is thus left out of this study. The dialect areas are presented in Figure 1.

The traditional division is based mostly on morphological and phonological features. The manual annotation of the messages is based on these same features, and it is carried out by the author, who holds a PhD in Finnish language with a special focus on language variation and change. An example of each dialect is presented in Table 1, with emphasis on dialectally marked features, and a more thorough presentation of the dialect features used in the annotation is presented in Appendix A.

There are messages written in the Helsinki slang, which does not appear in the traditional dialect division. The slang is a variety of Finnish characterized by Swedish loanwords. For example in Table 1, the words *luudata*, *gartsa*, *ookaamaan*, and *spora* are of Swedish origin (se. *loda*, *gata*, *åka*, *spårvagn*). The messages written in Helsinki slang are thus classified to their own group.

Many messages are written in a colloquial style, which is used especially by young speakers both in speech and online. The colloquial style is based on the Häme dialect, but it differs from it in some key features (such as the pronoun 'I', with *mä* in the colloquial style and *mää* in Häme). We thus include the colloquial style as a separate variety, but hypothesize that discerning these two varieties from each other will be difficult for the models.

2.1.2. Data split

The manually annotated data of each variety is split so that 90% of messages are used for training and 10% for testing. We create three different folds with random seeds of the train–test split to see if the methods are stable in their predictions. The resulting test sets are called **balanced** test sets.

The balanced test sets are the same for both the standard vs. non-standard messages and for dialectal differentiation. However, the labels are different for the two stages (standard or non-standard, the nine varieties), and the standard Finnish messages are excluded from the later stage.

We expect that the distribution of varieties in the full Suomi24 corpus is not balanced at all, with most messages being either in standard Finnish or in the colloquial style. Therefore, we take two more **random** test sets of 200 messages. The first set is used to evaluate the standard vs. non-standard division and the second random set is used to evaluate the identification of dialects. The distribution of messages in the manually annotated S24 dataset is presented in Table 2.

2.2. Additional Training Data

We create three different training datasets, each augmenting the previous:

1. using only the manually annotated S24 data (see Section 2.1.2),
2. adding data from the web (Murreviikko and Wikipedia), and
3. adding dialectal transcriptions (SKN).

The additional training data are introduced in more detail below, and sizes of the datasets in units and characters are presented in Table 3.

Murreviikko is a small corpus of tweets written in Finnish dialects between 2020–2022 (Kuparinen, 2023). The tweets have been annotated following the same dialect division as in this work (Itkonen, 1989) and normalized to standard Finnish. We use both the dialectal tweets and their normalizations.

Samples of Spoken Finnish (SKN) is a corpus of 99 interviews conducted in 50 Finnish-speaking municipalities (Institute for the Languages of Finland, 2021). The interviews have been transcribed on two levels of precision. We use the simpler transcriptions, which are written in the Finnish alphabet. We extract the interviewees' turns and keep only the ones which have at least 20 tokens, since shorter segments might not have enough dialectal features to be helpful for training.

Wikipedia We take a random sample of 50,000 paragraphs from the Finnish Wikipedia for discrimination between standard and non-standard Finnish. The data is extracted from a readily available Wikipedia collection from 2017 (Huovilainen,

CO	**ne on** tääl netis niitä peräkammarin poik**ii** , jotka ressukat ei **in real life** uskalla He ovat täällä netissä niitä peräkammarin poikia, jotka ressukat eivät oikeassa elämässä uskalla 'Those people online, living with their parents, are afraid in real life'
FN	Ei siinä mit**h**än pa**hh**aa ole vaikka murt**h**e**l**a pu**hh**uuki Ei siinä mitään pahaa ole vaikka murteella puhuukin 'There is nothing wrong in speaking with dialect'
HE	tottunu **luudaamaan** pitkin **gartsaa** ja **ookaamaan sporalla** ja **dösällä** tottunut kulkemaan pitkin katuja ja matkustamaan raitiovaunulla ja linja-autolla . 'used to walking around the streets and ride with trams and buses.'
HÄ	Täytyy vissiin tulla poikk**een siä** kylällä . täytyy vissiin tulla poikkeamaan siellä kylällä . 'I probably must stop by at the centre'
NO	Halua**kko**sää mait**ua** kah**a**viis ? Haluatko sinä maitoa kahviisi ? 'Do you want milk in your coffee?'
SA	To**e**vottavasti **outta** sel**e**vin**nä** p**iä**si**j**äisen vietosta Toivottavasti olette selvinneet pääsiäisen vietosta 'Hopefully you have survived Easter'
SE	**miul** tu**l** kutsumus to**n** toise**l** osasto**l** pi**t** käy**vv**ä lukasema**s** ku olin**kii** saa**nt** post**ii** minulle tuli kutsumus tuonne toiselle osastolle piti käydä lukaisemassa kun olinkin saanut postia 'I got drawn to the other department had to read when I got mail'
SO	Tämä sen täh**är**e **notta** ne ain**ua**t pyhähousut on kumminki vähä lyhk**ää**set Tämä sen tähden että ne ainoat pyhähousut ovat kuitenkin vähän lyhkäiset 'This because the only good pants are anyway too short'
SW	sill**o mää** e oi**kke** nää yhti**kkä** mi**ttä**n silloin minä en oikein näe yhtään mitään 'I don't see pretty much anything then'

Table 1: An example sequence from each variety in the manually annotated dataset (top), standard Finnish translation (middle) and English gloss (bottom). The marked features of each dialect are presented in bold. CO = Colloquial, FN = Far North, HE = Helsinki, HÄ = Häme, NO = Northern Ostrobothnia, SA = Savo, SE = South-East, SO = Southern Ostrobothnia, SW = South-West.

	Train	Test (bal.)	Test (rand.)	Total
St.	930	104	163	1197
Non-st.	2645	299	37	2981
CO	378	42	181	601
FN	271	31	2	304
HE	276	31	3	310
HÄ	226	26	3	255
NO	275	31	2	308
SA	306	34	2	342
SE	263	30	5	298
SO	362	41	2	405
SW	288	33	0	321

Table 2: Number of messages in the manually annotated dataset (S24) and their division to training and test sets. Test (bal.) = Balanced test set, Test (rand.) = Random test set. St. = Standard Finnish, Non-st. = Non-standard Finnish. Note that there are two random test sets: one for standard vs. non-standard, and one for the dialects.

	Web (u.)	Web (c.)	SKN (u.)	SKN (c.)
St.	50,344	26M	-	
FN	14	3155	284	269,777
HE	9	1572	-	-
HÄ	58	10,097	1738	992,924
NO	33	6506	375	246,656
SA	95	18,362	2225	1,396,737
SE	14	2848	815	490,396
SO	17	2848	336	228,580
SW	74	12,391	862	575,030

Table 3: Number of units (messages, utterances, paragraphs; u.) and characters (c.) in the external training data. Web = Murreviikko and Wikipedia, SKN = Samples of Spoken Finnish.

2019). As for SKN, we use a 20 token threshold for Wikipedia paragraphs.

3. Methods

3.1. Automatic Annotation Process

The messages posted on the Suomi24 forum are annotated in a three-step process. Firstly, they are divided based on the language used. The languages are identified with the HeLI-OTS tool (Jauhiainen et al., 2022), with which a small subset of the data had already been identified (City Digital Group, 2021b). According to the automatic classification, 90.8M messages are in Finnish (from a total of 93.7M messages). The Finnish messages include 4.5B tokens.

The messages deemed non-Finnish typically consist of non-sensical content, only URL's or long quotes in English in an otherwise Finnish text. Messages written entirely in another language are rare, but English, Swedish, Estonian, and Northern Sámi are the largest languages besides Finnish presented in the corpus.

The messages written in Finnish are divided to include either standard or non-standard language using the methods presented in Section 3.2. Finally, the same methods are used to classify the non-standard Finnish messages to the nine varieties, discussed in Section 2.1.1.

3.2. Dialect Identification Models

We experiment with five different models for the dialect identification tasks, as well as a majority vote ensemble. The models can be classified to three groups: traditional linear classification methods, retrained off-the-shelf tools and a fine-tuned neural model. The models are described in this order in the following.

SVM Support Vector Machines are one of the most used methodologies for language (and dialect) identification, given their easy implementation and good performance. An SVM-based method has been ranked first in several shared tasks focusing on discrimination between similar languages or dialects (Malmasi et al., 2016; Zampieri et al., 2017; Gaman et al., 2020). We train an SVM classifier on character n-grams ranging from 2 to 6, with tf-idf weighting.

NB Naïve Bayes is a standard tool used in language identification and other classification tasks. We train an NB classifier on the same range of character n-grams and weighting as the SVM classifier. Both SVM and NB are implemented with the Python library *scikit-learn* (Pedregosa et al., 2011).

fastText is an off-the-shelf tool trained for language identification (Joulin et al., 2016).[2] The model uses a bag of words (and n-grams) approach, providing very fast training. We train the model for 25 epochs with our datasets with a learning rate of 1.0.

HeLI is an off-the-shelf tool trained for language identification (Jauhiainen et al., 2022).[3] It adopts a Naïve Bayes based approach with word and character n-gram models used in a fall-back system. An earlier version of the method has fared well in the shared tasks on discrimination between similar languages (Jauhiainen et al., 2017, 2018).

FinBERT is the Finnish version of the BERT model (Devlin et al., 2019) trained on mostly Web-based data (Virtanen et al., 2019). It is worth noting that the available Suomi24 messages were used for training of the original model. We fine-tune the model[4] for the classification task for 5 epochs.

Ensemble is a majority vote on the predictions given by the aforementioned models. If some models are found to underperform greatly, they are left out of the vote.

3.3. Evaluation

We evaluate the performance of the models with the weighted F1-score. The weighted score takes the label imbalance in the random test sets into account. The corresponding precision and recall scores are presented in Appendix B. The scores are averages over three folds and we also present the standard deviations.

4. Results

4.1. Dialect Identification

We report the performance of the models presented in Section 3.2 on the two test sets with the weighted F1-score. The results are presented in two stages, first for the standard vs. non-standard Finnish division, and thereafter for the dialect identification.

4.1.1. Standard vs. Non-standard Finnish

The scores for the different training sets introduced in Section 2.2 and the two test sets are presented in Table 4. From the single models, SVM achieves the best scores in both test sets. This was to be expected, as SVM-based approaches have fared well

[2] https://fasttext.cc/

[3] https://github.com/tosaja/HeLI

[4] Fine-tuning with Simple Transformers https://simpletransformers.ai/.

Model	S24	+ Web	+ SKN
Balanced			
NB	$0.45^{\pm 0.01}$	$0.25^{\pm 0.01}$	$0.59^{\pm 0.01}$
SVM	$0.89^{\pm 0.02}$	$0.90^{\pm 0.01}$	$\mathbf{0.91}^{\pm 0.02}$
fastText	$0.87^{\pm 0.01}$	$0.86^{\pm 0.02}$	$0.87^{\pm 0.01}$
HeLI	$0.86^{\pm 0.02}$	$0.59^{\pm 0.01}$	$0.76^{\pm 0.03}$
FinBERT	$0.76^{\pm 0.23}$	$0.21^{\pm 0.00}$	$0.21^{\pm 0.00}$
Ensemble	$0.89^{\pm 0.00}$	$0.89^{\pm 0.02}$	$0.89^{\pm 0.02}$
Random sample			
NB	$0.06^{\pm 0.00}$	$0.73^{\pm 0.00}$	$0.75^{\pm 0.01}$
SVM	$0.72^{\pm 0.02}$	$\mathbf{0.86}^{\pm 0.01}$	$\mathbf{0.86}^{\pm 0.00}$
fastText	$0.77^{\pm 0.02}$	$0.85^{\pm 0.00}$	$0.85^{\pm 0.01}$
HeLI	$0.80^{\pm 0.02}$	$0.37^{\pm 0.02}$	$0.62^{\pm 0.01}$
FinBERT	$0.58^{\pm 0.37}$	$0.73^{\pm 0.00}$	$0.73^{\pm 0.00}$
Ensemble	$0.73^{\pm 0.02}$	$0.85^{\pm 0.00}$	$\mathbf{0.86}^{\pm 0.01}$

Table 4: Weighted F1 scores (↑) on the standard vs. non-standard testsets. S24 = Manual Suomi24 annotation. Web = Finnish Wikipedia and Murreviikko tweets. SKN = Samples of Spoken Finnish transcripts. NB = Naïve Bayes. SVM = Support Vector Machine. Ensemble = A majority vote of SVM, fastText and HeLI.

Model	S24	+ Web	+ SKN
Balanced			
NB	$0.56^{\pm 0.01}$	$0.55^{\pm 0.01}$	$0.10^{\pm 0.00}$
SVM	$0.80^{\pm 0.01}$	$0.81^{\pm 0.01}$	$0.78^{\pm 0.01}$
FastText	$0.74^{\pm 0.02}$	$0.75^{\pm 0.02}$	$0.70^{\pm 0.00}$
HeLI	$0.68^{\pm 0.02}$	$0.68^{\pm 0.02}$	$0.63^{\pm 0.03}$
FinBERT	$0.81^{\pm 0.01}$	$0.81^{\pm 0.01}$	$0.78^{\pm 0.01}$
Ensemble	$\mathbf{0.82}^{\pm 0.01}$	$\mathbf{0.82}^{\pm 0.01}$	$0.78^{\pm 0.00}$
Random sample			
NB	$0.86^{\pm 0.01}$	$0.79^{\pm 0.02}$	$0.00^{\pm 0.00}$
SVM	$0.84^{\pm 0.01}$	$0.83^{\pm 0.01}$	$0.74^{\pm 0.00}$
FastText	$0.76^{\pm 0.01}$	$0.76^{\pm 0.01}$	$0.71^{\pm 0.01}$
HeLI	$0.72^{\pm 0.01}$	$0.73^{\pm 0.01}$	$0.80^{\pm 0.01}$
FinBERT	$0.85^{\pm 0.02}$	$\mathbf{0.87}^{\pm 0.01}$	$0.82^{\pm 0.02}$
Ensemble	$0.85^{\pm 0.01}$	$0.85^{\pm 0.01}$	$0.81^{\pm 0.00}$

Table 5: Weighted F1 scores (↑) on the dialect testsets. S24 = Suomi24. Web = Murreviikko tweets. SKN = Samples of Spoken Finnish transcripts. NB = Naïve Bayes. SVM = Support Vector Machine. Ensemble = A majority vote of SVM, fastText, HeLI and FinBERT.

in previous dialect identification tasks. The addition of training data seems to enhance the performance of the SVM method especially in the random test set.

The re-trained fastText tool achieves very similar scores to the SVM, and given its much faster performance, it could be a potential solution as well. The Naïve Bayes based approaches (NB and HeLI) interestingly turn worse with the addition of training data.

Finally, the neural FinBERT solution provides the best individual run with the S24 data (weighted F1-score of 0.94), but it is very unstable which is reflected in the standard deviation. With additional data, the FinBERT model predicts only one label. The NB method also predicts just one label in the random sample with additional data.

The ensemble is a majority vote of the SVM's, fastText's and HeLI's predictions. FinBert and NB are excluded from the vote due to their inconsistent performance. The ensemble achieves very similar scores to the best-performing SVM model.

We use the SVM model with all of the training data to classify the messages to either standard Finnish or non-standard Finnish. There are 15.9 million non-standard messages, which are used for the final dialect identification.

4.1.2. Classification of Dialects

The weighted F1-scores for the dialect identification task are presented in Table 5. The SVM method is still very stable across training datasets. The fastText method is significantly worse in this task than in the previous one. The NB method is still unstable, with some high scores and a total mislabeling in the random test set with all of the training data. The Naïve Bayes based HeLI, however, is more stable in this task than in Section 4.1.1.

The neural FinBERT provides the best score for the random test set with the additional Web-data and is close to the best score on the balanced sample as well. The high standard deviations have also disappeared. The ensemble method is close second in scores, offering the best results in the balanced set. We use the FinBERT method for the final labeling of varieties.

Interestingly, the addition of the dialect transcriptions seems to hurt all models except HeLI. This could indicate that the written dialects are somewhat different from the spoken and transcribed dialects. It is also possible that the themes discussed in the spoken dialect interviews are too different (mostly agriculture in rural Finland in the 1960s) from the themes discussed in the Internet forum. This effect was not visible in the standard vs. non-standard division, since the models only picked up on the non-standard structure of the text. Moreover, as the dialect transcriptions do not contain the colloquial style, the SKN training data might in fact be harmful for the recognition of this particular variety.

After labeling the messages with the FinBERT model using S24+Web training data, we see that our random sample of non-standard texts repre-

Total	Finnish	Non-standard	Dialectal
93.7M	90.8M	15.9M	3.5M

Table 6: Messages in the Suomi24 corpus in total, in Finnish (labeled with HeLI-OTS), in non-standard Finnish (with SVM on all of the training data) and in the dialects (with FinBERT on S24+Web training data).

sents the whole dataset: of the 15.9M non-standard messages, 12.4M are written in the colloquial style. Thus, only 3.5M messages are written using the other eight varieties. The total number of messages in each stage of the classification are presented in Table 6.

4.2. Error Analysis

We analyze the errors produced by the best model, FinBERT fine-tuned with the additional Web data, on the balanced test set. The confusion matrix for the predictions and true labels is presented in Figure 2.

As expected, the colloquial style and Häme dialect are most often confused with one another. Other mislabelings are also easily explainable: the Southern Ostrobothnian dialect and the Häme dialect share a lot of features, as do the Northern Ostrobothnia and Savo.

All in all, it is apparent that the Häme dialect is predicted much more often than it actually appears. The dialect area is in a central location in Finland and is characterized by many features that also appear in other dialects and the colloquial style, which makes it difficult to discern from others.

4.3. Change Over Time

Given the longitudinal nature of the Suomi24 corpus, it is possible to study the changes in language use over time. We focus on the change between the different dialects. Since the colloquial style is by far the most popular non-standard variety, we compare its proportion of messages to all the other dialects in Figure 3.

The proportion of messages written in the colloquial style has been in decline for most of the forum's lifetime, with a small increase around 2011. It is possible that the new social media platforms introduced in the 2000s and 2010s (e.g., Facebook, Twitter, Instagram) have had an impact on the volume of non-standard messages, but there are no sudden drops. The decline in the usage of the colloquial style can nonetheless be interpreted so that young users were not using the forum as much in the 2010's as they did in the 2000's.

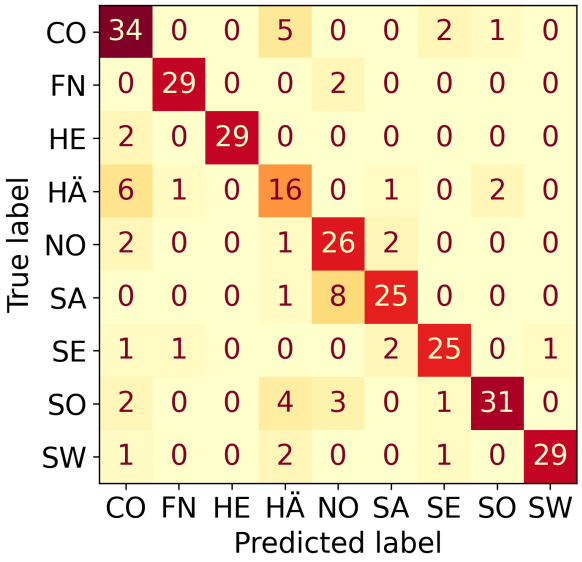

Figure 2: A confusion matrix of the predicted and true labels. Predictions are produced by the FinBERT model on the balanced test set.

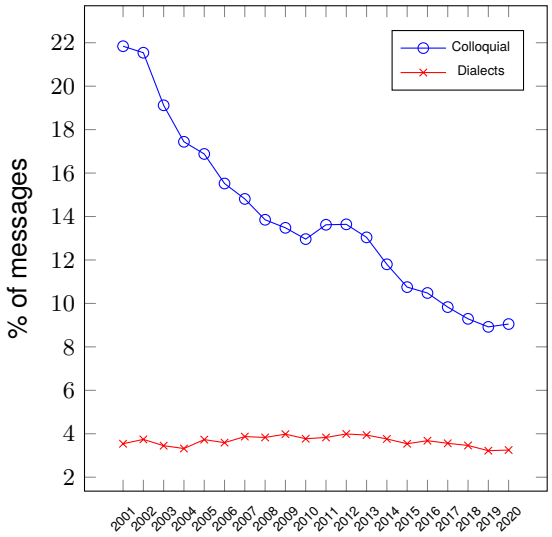

Figure 3: Proportion of messages written in the colloquial style and in the dialects in the Suomi24 forum over time.

While the proportion of messages written in the colloquial style has decreased, the proportion of the traditional dialects has stayed relatively stable at around 3.5%. This indicates that there has been a small user-base writing in dialects for the forum's life-span, and it has not been affected by the major changes that have happened in computer-mediated communication such as the introduction of smartphones or new social media platforms.

Since the proportion of the colloquial style compared to the dialects in Figure 3 is much higher, we take a further look at the change of other dialects alone in Figure 4. The Häme dialect is the

most popular of the other dialects. However, this dialect was often mislabeled in Figure 2, and it could indicate that some of the messages are in fact wrongly identified. The usage of the dialect is also in slow decline, which further shows that some colloquial messages are possibly labeled as the Häme dialect.

The Northern Ostrobothnian dialect is a similar case, as it also appears more often than the other dialects, and also had a lot of mislabelings. It is easily confused with Savo, Southern Ostrobothnia and Far North, as many Northern Ostrobothnian features also appear in these dialects. All other dialects are relatively rare and the change patterns are rather similar.

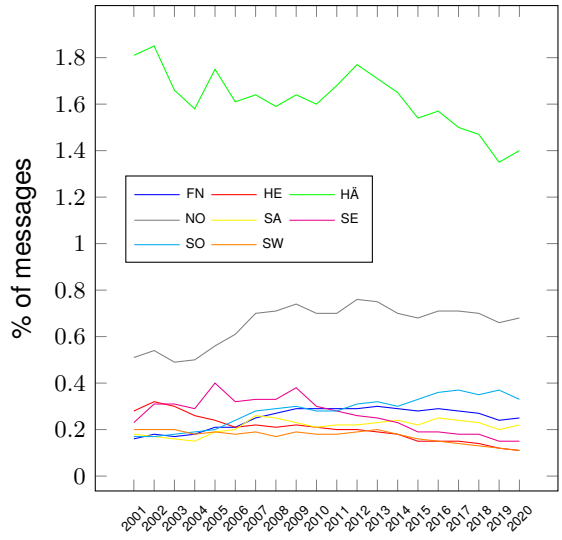

Figure 4: Proportion of messages written in dialects in the forum over time. Colloquial style omitted.

5. Conclusion

In this paper, we have presented Murre24, a collection of dialectal messages identified from the largest Finnish Internet forum, Suomi24. We manually annotated a set of messages from the corpus and used that dataset, along with additional training data, to build five different dialect identification models and a majority vote ensemble. All 94M messages of the corpus were annotated automatically in a three-step process: first excluding non-Finnish content, then excluding standard Finnish, and finally identifying the dialect used, utilizing a division to nine varieties.

In the stage of differentiating standard and non-standard Finnish, a traditional linear SVM model offered best performance. It remained stable across different training sets when working on the balanced dataset, but the model with the most training data (Web + SKN) achieved best scores on the randomly selected sample. The neural FinBERT model was the best on individual runs, but turned out to be very unstable over three runs. There were 15.9M non-standard messages in the corpus according to our SVM model.

The non-standard messages were classified to nine varieties with the same five methods and a majority vote ensemble. In this latter stage, the FinBERT model remained stable and offered the best overall performance. In both stages, the Naïve Bayes based approaches fared poorly. When labeling messages with the FinBERT model trained on the additional Web data, there were 3.5M messages in the traditional dialects and 12.4M messages in the colloquial style.

We found that the proportion of messages written in the colloquial style has been in steady decline for the life-span of the forum, but the usage of other dialects has stayed relatively stable. A further analysis of errors produced also showed that the dialects of Häme and Northern Ostrobothnia were most often mislabeled due to shared features with other dialects.

6. Limitations

Classifying messages to pre-existing groups such as traditional dialects enforces old categorizations and does not give space to, for instance, mixing of dialects or new varieties, thus possibly preventing new scientific findings. An unsupervised or generative method, such as the latent Dirichlet allocation (Blei et al., 2003), could provide more insights into the variation encountered in the data without pre-existing class labels.

The decision to keep the Häme dialect and the colloquial style separate could be critically discussed. The colloquial style is mostly based on the Häme dialect, even though it differs from it in some aspects. The idea behind separating the varieties was to provide a more sensible split to traditional dialects (including Häme) and the contemporary colloquial style. For instance, dialectologists could be interested in the traditional dialects and not as much in the colloquial style (and vice versa for sociolinguists).

Finally, there are dialectally ambiguous messages, either because of short length with no clear dialectal markers or because of contradicting dialectal features. Labeling such messages is difficult even for a human annotator. A possible solution for the issue would be multilabel classification, where a message can have different labels instead of just one.

7. Ethical considerations

As with any social media outlet, the Suomi24 forum has its issues with hate speech, racism, and

misogyny, and one can find such content also in the dialectal messages. In the manual annotation, such messages were excluded. In the automatic annotation, however, this was not possible, and the dialectally annotated collection could include messages with harmful content. Future work could identify such messages in the forum corpus, regardless of the used language variety.

8. Acknowledgements

This work has been supported by the Academy of Finland through project No. 342859 "CorCoDial – Corpus-based computational dialectology" and by the Kone Foundation through project "LANGAWARE".

9. Bibliographical References

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Mohamed Ali. 2018. Character level convolutional neural network for German dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 172–177, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Maha J. Althobaiti. 2022. Creation of annotated country-level dialectal arabic resources: An unsupervised approach. *Natural Language Engineering*, 28(5):607–648.

Jeremy Barnes, Petter Mæhlum, and Samia Touileb. 2021. NorDial: A preliminary corpus of written Norwegian dialect use. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 445–451, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Léger. 2019. Improving cuneiform language identification with BERT. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 17–25, Ann Arbor, Michigan. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Giacomo Camposampiero, Quynh Anh Nguyen, and Francesco Di Stefano. 2022. The curious case of logistic regression for Italian languages and dialects identification. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 86–98, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding u.s. regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.

Terho Itkonen. 1989. *Nurmijärven murrekirja*. Suomalaisen Kirjallisuuden Seuran toimituksia ; 498. Suomalaisen kirjallisuuden seura, Helsinki.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018. HeLI-based experiments in Swiss German dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 254–262, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022. HeLI-OTS, off-the-shelf language identifier for text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3912–3922, Marseille, France. European Language Resources Association.

Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen, and Krister Lindén. 2020. Uralic language identification (ULI) 2020 shared task dataset and the wanca 2017 corpora. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 173–185, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2017. Evaluating HeLI with non-linear mappings. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 102–108, Valencia, Spain. Association for Computational Linguistics.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019a. Discriminating between Mandarin Chinese and Swiss-German varieties using adaptive language models. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 178–187, Ann Arbor, Michigan. Association for Computational Linguistics.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019b. Automatic language identification in texts: A survey. *J. Artif. Int. Res.*, 65(1):675–682.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Olli Kuparinen. 2023. Murreviikko - a dialectologically annotated and normalized dataset of Finnish tweets. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 31–39, Dubrovnik, Croatia. Association for Computational Linguistics.

Nikola Ljubešić, Tanja Samardžić, and Curdin Derungs. 2016. TweetGeo - a tool for collecting, processing and analysing geo-encoded linguistic data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3412–3421, Osaka, Japan. The COLING 2016 Organizing Committee.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Paul McNamee. 2005. Language identification: A solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101.

Hamdy Mubarak. 2018. Converting Dialectal Arabic to Modern Standard Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ahmed Salama, Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2014. Youdacc: the youtube dialectal arabic comment corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Simone Ueberwasser and Elisabeth Stark. 2017. What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*, 5(84).

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip

Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *ArXiv*, abs/1912.07076.

Nianheng Wu, Eric DeMattos, Kwok Him So, Pinzhen Chen, and Çağrı Çöltekin. 2019. Language discrimination and transfer learning for similar languages: Experiments with feature combinations and adaptation. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63, Ann Arbor, Michigan. Association for Computational Linguistics.

George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. 2020. Exploring the power of Romanian BERT for dialect identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 232–241, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626.

10. Language Resource References

City Digital Group. 2021a. The Suomi 24 Corpus 2001-2020, VRT version.

City Digital Group. 2021b. The Suomi24 Corpus 2018-2020, VRT version.

Tatu Huovilainen. 2019. Finnish Wikipedia 2017, source.

Institute for the Languages of Finland. 2021. Samples of Spoken Finnish, VRT Version.

A. Dialect Features

This section presents a collection of the dialectal features used as a basis for the annotation and the searches for dialectal messages. Vowels experiencing vowel harmony are in capital letters (A = *a* or *ä* depending on the vowel harmony). The standard Finnish alternatives are in brackets.

Colloquial

- Personal pronouns *mä* 'I' and *sä* 'you' (minä, sinä)
- Monophthongization of all A-final vowel combinations: *kahvii* 'coffee-PT', *korkee* 'high' (kahvia, korkea)
- Deletion of *d* after *h* in high-frequency lexemes: *tehä* 'to do', *kahelta* 'at two' (tehdä, kahdelta)
- Apocope especially after *l*: *talol* 'at/to the house', *viel* 'still', *kyl* 'yes' (talolla, vielä, kyllä)
- English words, e.g., *for real*

Far North

- Personal pronouns *mie* 'I' *met, tet, het* 'we, you, they' (minä, me, te, he)
- Metathesis of historical *h*: *menhän saunhan* 'let's go to the sauna' (mennään saunaan)
- Consonant gemination: *tekkee* 'does' (tekee)
- Deletion of *d*: *viien* 'five's' (viiden)

Helsinki

- Swedish loanwords, e.g., *gartsa* 'road', *stadi* 'city, Helsinki' (katu, kaupunki)

Häme

- Substitution of *d* with *r* or *l*: *meirän, meilän* 'ours' (meidän)
- Personal pronouns *mää* 'I' and *sää* 'you' (minä, sinä)
- Diphthong opening: *nuari* 'young' (nuori)
- No consonant gradation in *nk*: *kenkät* 'shoes' (kengät)
- Deletion of -mA in -mA-infinitive: *tuu poikkeen* 'come and stop by' (tule poikkeamaan)
- Monophthongization of A-final vowel combinations in OA and eA: *palloo* 'ball-PT', *korkee* 'high' (palloa, korkea)

Northern Ostrobothnia

- Epenthetic vowels: *vanaha* 'old' (vanha)

- Consonant gemination: *tekkee* 'does' (tekee)
- *UA* instead of *OA*: *sanua* 'to say' (sanoa)
- Personal pronoun *nää* 'you' (sinä)

Savo

- Epenthetic vowels: *vanaha* 'old' (vanha)
- Vast consonant gemination: *tekkee* 'does' *syömmään* '(go) eat' (tekee, syömään)
- Diphthongization: *piä* 'head' (pää)
- Diphthong reduction: *koera* 'dog' (koira)
- Personal pronouns *myö, työ, hyö* 'we, you, they' (me, te, he)

South-East

- Personal pronouns *myö, työ, hyö* 'we, you, they' (me, te, he)
- Personal pronouns *mie, sie, hää* 'I, you, he/she' (minä, sinä, hän)
- Vast apocope: *talolt* 'from the house', *talos* 'in the house' (talolta, talossa)
- Clitic *-kii* 'as well': *siekii* 'you too' (sinäkin)
- *-nt* as participle perfect: *saant* 'got' (saanut)

Southern Ostrobothnia

- Preserved historical *h*: *saunahan* 'to the sauna' (saunaan)
- Monophthongization of i-final diphthongs: *punaanen* 'red' (punainen)
- Substitution of *d* with *r*: *meirän* 'ours' (meidän)
- Epenthetic vowels: *vanaha* 'old' (vanha)
- Inessive case marker *-hnA*: *mihnä* 'where' (missä)
- *UA* instead of *OA*: *sanua* 'to say' (sanoa)
- Words *jotta* and *notta* instead of *että* 'that'

South-West

- Special consonant gemination affecting k, p, t and s: *jokke* 'to the river' (jokeen)
- Vast apocope: *talolt* 'from the house', *talos* 'in the house' (talolta, talossa)
- Syncope: *suamlase* 'Finnish people' (suomalaiset)
- Diphthong opening: *nuar* 'young' (nuori)
- Substitution of *d* with *r*: *meirä* 'ours' (meidän)
- Shortening of long vowels in all syllables after the first: *kala* 'fish-PT' (kalaa)
- Loss of final consonants: *kala* 'fish-GEN' (kalan)

Model	S24	+ Web	+ SKN
	Balanced		
NB	$0.87^{\pm 0.00}$	$0.63^{\pm 0.00}$	$0.68^{\pm 0.01}$
SVM	$0.91^{\pm 0.01}$	$0.89^{\pm 0.01}$	$0.89^{\pm 0.02}$
FastText	$0.88^{\pm 0.00}$	$0.85^{\pm 0.02}$	$0.85^{\pm 0.01}$
HeLI	$0.85^{\pm 0.02}$	$0.81^{\pm 0.03}$	$0.81^{\pm 0.04}$
FinBERT	$0.74^{\pm 0.26}$	$0.13^{\pm 0.00}$	$0.13^{\pm 0.00}$
Ensemble	$0.90^{\pm 0.01}$	$0.89^{\pm 0.02}$	$0.89^{\pm 0.02}$
	Random sample		
NB	$0.03^{\pm 0.00}$	$0.66^{\pm 0.00}$	$0.85^{\pm 0.00}$
SVM	$0.87^{\pm 0.00}$	$0.86^{\pm 0.01}$	$0.86^{\pm 0.00}$
FastText	$0.85^{\pm 0.01}$	$0.86^{\pm 0.01}$	$0.87^{\pm 0.01}$
HeLI	$0.85^{\pm 0.00}$	$0.84^{\pm 0.00}$	$0.83^{\pm 0.00}$
FinBERT	$0.61^{\pm 0.41}$	$0.66^{\pm 0.00}$	$0.66^{\pm 0.00}$
Ensemble	$0.86^{\pm 0.01}$	$0.86^{\pm 0.01}$	$0.87^{\pm 0.01}$

Table 7: Precision scores (↑) on the standard and non-standard testsets. S24 = Suomi24. Web = Murreviikko tweets. SKN = Samples of Spoken Finnish transcripts. NB = Naïve Bayes. SVM = Support Vector Machine. Ensemble = A majority vote of SVM, fastText, and HeLI.

Model	S24	+ Web	+ SKN
	Balanced		
NB	$0.51^{\pm 0.01}$	$0.52^{\pm 0.01}$	$0.72^{\pm 0.01}$
SVM	$0.87^{\pm 0.02}$	$0.92^{\pm 0.01}$	$0.92^{\pm 0.02}$
FastText	$0.87^{\pm 0.01}$	$0.89^{\pm 0.02}$	$0.88^{\pm 0.01}$
HeLI	$0.87^{\pm 0.02}$	$0.59^{\pm 0.01}$	$0.73^{\pm 0.03}$
FinBERT	$0.78^{\pm 0.20}$	$0.50^{\pm 0.00}$	$0.50^{\pm 0.00}$
Ensemble	$0.88^{\pm 0.00}$	$0.90^{\pm 0.02}$	$0.90^{\pm 0.02}$
	Random sample		
NB	$0.19^{\pm 0.00}$	$0.82^{\pm 0.00}$	$0.82^{\pm 0.00}$
SVM	$0.68^{\pm 0.02}$	$0.86^{\pm 0.01}$	$0.85^{\pm 0.00}$
FastText	$0.75^{\pm 0.02}$	$0.84^{\pm 0.00}$	$0.84^{\pm 0.01}$
HeLI	$0.79^{\pm 0.02}$	$0.37^{\pm 0.02}$	$0.58^{\pm 0.01}$
FinBERT	$0.62^{\pm 0.31}$	$0.82^{\pm 0.00}$	$0.82^{\pm 0.00}$
Ensemble	$0.69^{\pm 0.02}$	$0.85^{\pm 0.00}$	$0.85^{\pm 0.01}$

Table 8: Recall scores (↑) on the standard and non-standard testsets. S24 = Suomi24. Web = Murreviikko tweets. SKN = Samples of Spoken Finnish transcripts. NB = Naïve Bayes. SVM = Support Vector Machine. Ensemble = A majority vote of SVM, fastText, and HeLI.

B. Precision and Recall

The precision scores for the standard vs. non-standard classification are presented in Table 7 and recall scores in Table 8. For the final dialect classification, the precision scores are presented in Table 9 and recall scores in Table 10.

Model	S24	+ Web	+ SKN
	Balanced		
NB	$0.72^{\pm 0.01}$	$0.71^{\pm 0.00}$	$0.29^{\pm 0.06}$
SVM	$0.81^{\pm 0.01}$	$0.82^{\pm 0.00}$	$0.79^{\pm 0.00}$
FastText	$0.75^{\pm 0.02}$	$0.76^{\pm 0.01}$	$0.71^{\pm 0.00}$
HeLI	$0.72^{\pm 0.01}$	$0.72^{\pm 0.02}$	$0.72^{\pm 0.02}$
FinBERT	$0.81^{\pm 0.02}$	$0.82^{\pm 0.01}$	$0.79^{\pm 0.00}$
Ensemble	$0.83^{\pm 0.01}$	$0.84^{\pm 0.01}$	$0.79^{\pm 0.00}$
	Random sample		
NB	$0.85^{\pm 0.00}$	$0.84^{\pm 0.00}$	$0.00^{\pm 0.00}$
SVM	$0.89^{\pm 0.00}$	$0.89^{\pm 0.01}$	$0.88^{\pm 0.00}$
FastText	$0.86^{\pm 0.02}$	$0.88^{\pm 0.01}$	$0.87^{\pm 0.00}$
HeLI	$0.86^{\pm 0.00}$	$0.86^{\pm 0.01}$	$0.86^{\pm 0.00}$
FinBERT	$0.91^{\pm 0.01}$	$0.91^{\pm 0.01}$	$0.90^{\pm 0.01}$
Ensemble	$0.89^{\pm 0.01}$	$0.89^{\pm 0.00}$	$0.88^{\pm 0.00}$

Table 9: Precision scores (↑) on the dialect testsets. S24 = Suomi24. Web = Murreviikko tweets. SKN = Samples of Spoken Finnish transcripts. NB = Naïve Bayes. SVM = Support Vector Machine. Ensemble = A majority vote of SVM, fastText, HeLI and FinBert.

Model	S24	+ Web	+ SKN
	Balanced		
NB	$0.59^{\pm 0.00}$	$0.59^{\pm 0.02}$	$0.17^{\pm 0.00}$
SVM	$0.80^{\pm 0.01}$	$0.81^{\pm 0.01}$	$0.78^{\pm 0.01}$
FastText	$0.74^{\pm 0.02}$	$0.74^{\pm 0.02}$	$0.69^{\pm 0.00}$
HeLI	$0.67^{\pm 0.02}$	$0.67^{\pm 0.02}$	$0.62^{\pm 0.02}$
FinBERT	$0.80^{\pm 0.01}$	$0.81^{\pm 0.01}$	$0.78^{\pm 0.01}$
Ensemble	$0.82^{\pm 0.01}$	$0.82^{\pm 0.01}$	$0.78^{\pm 0.00}$
	Random sample		
NB	$0.88^{\pm 0.01}$	$0.75^{\pm 0.03}$	$0.01^{\pm 0.00}$
SVM	$0.81^{\pm 0.01}$	$0.79^{\pm 0.01}$	$0.66^{\pm 0.00}$
FastText	$0.70^{\pm 0.02}$	$0.69^{\pm 0.01}$	$0.62^{\pm 0.01}$
HeLI	$0.64^{\pm 0.02}$	$0.65^{\pm 0.01}$	$0.76^{\pm 0.01}$
FinBERT	$0.82^{\pm 0.02}$	$0.84^{\pm 0.01}$	$0.78^{\pm 0.03}$
Ensemble	$0.82^{\pm 0.01}$	$0.83^{\pm 0.01}$	$0.77^{\pm 0.00}$

Table 10: Recall scores (↑) on the dialect testsets. S24 = Suomi24. Web = Murreviikko tweets. SKN = Samples of Spoken Finnish transcripts. NB = Naïve Bayes. SVM = Support Vector Machine. Ensemble = A majority vote of SVM, fastText, HeLI and FinBert.