Findings of the VarDial Evaluation Campaign 2021

Bharathi Raja Chakravarthi[1], Mihaela Găman[2], Radu Tudor Ionescu[2], Heidi Jauhiainen[3], Tommi Jauhiainen[3,9], Krister Lindén[3], Nikola Ljubešić[4,5], Niko Partanen[3], Ruba Priyadharshini[6], Christoph Purschke[7], Eswari Rajagopal[8] Yves Scherrer[3], Marcos Zampieri[9]

[1]National University of Ireland Galway, [2]University of Bucharest, [3]University of Helsinki, [4]Jožef Stefan Institute, [5]University of Ljubljana, [6]Madurai Kamaraj University, [7]University of Luxembourg, [8]National Institute of Technology Tiruchirappalli, [9]Rochester Institute of Technology

vardialworkshop@gmail.com

Abstract

This paper describes the results of the shared tasks organized as part of the VarDial Evaluation Campaign 2021. The campaign was part of the eighth workshop on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects (VarDial), co-located with EACL 2021. Four separate shared tasks were included this year: Dravidian Language Identification (DLI), Romanian Dialect Identification (RDI), Social Media Variety Geolocation (SMG), and Uralic Language Identification (ULI). DLI was organized for the first time and the other three continued a series of tasks from previous evaluation campaigns.

1 Introduction

The computational processing of similar languages, varieties and dialects is a vibrant area of research discussed in a recent survey (Zampieri et al., 2020). Co-located with international conferences, the workshop series on NLP for Similar Languages, Varieties and Dialects (VarDial) has become the main workshop on this topic reaching its eighth edition in 2021. Since its first edition, VarDial has included well-attended shared tasks on topics such as language and dialect identification, morphosyntactic tagging, and cross-lingual parsing. These shared tasks became part of the VarDial Evaluation Campaigns featuring multiple shared tasks organized yearly with the workshop (Zampieri et al., 2017, 2018, 2019; Găman et al., 2020).

Together with VarDial 2021, we organized the fifth edition of the VarDial Evaluation Campaign.[1] The VarDial Evaluation Campaign 2021 featured four shared tasks addressing different aspects of language and dialect identification. In this paper, we present the results and main findings of these shared tasks. Section 4 presents the Dravidian Language Identification (DLI) shared task included for the first time at VarDial 2021. The Romanian Dialect Identification (RDI) shared task is described in Section 5 and the Social Media Variety Geolocation (SMG) task is presented in Section 6. These two tasks are task re-runs from VarDial 2020 with augmented datasets prepared for VarDial 2021. Finally, the Uralic Language Identification (ULI) shared task, described in Section 7, is an open leaderboard shared task that ran between VarDial 2020 and 2021. We include references to the 8 system description papers written by the participants of the campaign in Table 1.

2 Shared Tasks at VarDial 2021

Dravidian Language Identification (DLI): Dravidian languages are a language family spoken mainly in the south of India (Chakravarthi, 2020). The four major literary Dravidian languages are Tamil (ISO 639-3: tam), Telugu (ISO 639-3: tel), Malayalam (ISO 639-3: mal), and Kannada (ISO 639-3: kan). Tamil, Malayalam, and Kannada are closely related belonging to the Tamil-Kannada subgroup. All three languages have official status in the Government of India. Outside India, Tamil also has official status in Sri Lanka and Singapore. These languages are widely considered to be under-resourced (Thavareesan and Mahesan, 2019, 2020a,b). The DLI shared task provides participants with a collection of 16,672 YouTube comments as training set. The comments contain code-mixed sentences with English and one of the South Dravidian languages (Tamil, Malayalam or Kannada). All comments were written in the Latin script (Non-native script). The task is to identify the language of each comment.

[1] https://sites.google.com/view/vardial2021/evaluation-campaign

Proceedings of the 8th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 1–11

Romanian Dialect Identification (RDI): The 2021 Romanian Dialect Identification shared task is at the third iteration, following the 2019 Moldavian vs. Romanian Cross-Dialect Topic identification (MRC) (Zampieri et al., 2019) and the 2020 Romanian Dialect Identification (RDI) (Găman et al., 2020) shared tasks. The 2021 RDI shared task is formulated as a cross-domain binary classification by dialect problem, in which a classification model is required to discriminate between the Moldavian (MD) and the Romanian (RO) subdialects. This year, we provided participants with an augmented version of the MOROCO data set (Butnaru and Ionescu, 2019) for training, which contains Moldavian and Romanian samples of text collected from the news domain. Last year's test set of tweets (Găman and Ionescu, 2020b) is used for validation. A new set of tweets has been collected for the 2021 shared task. The task has two formats, open and closed. In the closed format, participants are not allowed to use external data to train their models. In the open format, participants are allowed to use external resources such as unlabeled corpora, lexicons and pre-trained embeddings (e.g. BERT), but the use of additional labeled data is still not allowed.

Social Media Variety Geolocation (SMG): In contrast to most past and present VarDial tasks, the SMG task is framed as a geolocation task: given a text, the participants have to predict its geographic location in terms of latitude and longitude coordinates. This setup addresses the common issue that defining a set of discrete labels is not trivial for many language areas where there is a continuum between varieties rather than clear-cut borders. The SMG task is split into three subtasks covering different language areas: the **BCMS** subtask is focused on geolocated tweets published in the area of Croatia, Bosnia and Herzegovina, Montenegro and Serbia in the Serbo-Croatian (HBS) macrolanguage (Ljubešić et al., 2016); the **DE-AT** subtask focuses on Jodel conversations initiated in Germany and Austria, which are written in standard German but commonly contain regional and dialectal forms; the **CH** subtask is based on Jodel conversations initiated in Switzerland, which were found to be held majoritarily in Swiss German dialects (Hovy and Purschke, 2018). All three subtasks used the same data format and evaluation methodology.

Uralic Language Identification (ULI): This task focuses on discriminating between the languages in the Uralic group as defined by the ISO 639-3 standard. Following VarDial 2020, ULI 2021 was an open public leaderboard competition where participants were able to submit at any point until the final submission date. A leaderboard page was set up to inform the participants of the current high scores and as a way to get more detailed information.[2] The task included 29 individual relevant languages, some of which are very closely related, such as Erzya (myv) and Moskha (mdf), or Livvi (olo) and Ludian (lud). The languages are spoken in Scandinavia, Estonia and Finland, and within the Russian Federation in a region that extends far into Siberia. In addition to the relevant languages, the task featured 149 non-relevant languages.

3 Participating Teams

A total of nine teams submitted runs to one or more shared tasks in this year's VarDial evaluation campaign. In Table 1, we list the teams that participated in the shared tasks, including references to the 8 system description papers which will be published as parts of the VarDial workshop proceedings. Detailed information about the submissions in each respective task is included in the following sections of this report.

4 Dravidian Language Identification (DLI)

4.1 Dataset

The DLI task is based on three datasets from YouTube comments (Chakravarthi et al., 2020b,a; Hande et al., 2020). In the 2021 (DLI) shared task, participants have to train a model on comments written in Roman script. Our corpora contains all the three types of code-mixed sentences: Inter-Sentential switch, Intra-Sentential switch and Tag switching. All comments were written in Roman script (Non-native script) with either one of the south Dravidian (Tamil, Malayalam, and Kannada) grammar with English lexicon or English grammar with south Dravidian lexicons (Jose et al., 2020; Priyadharshini et al., 2020). The comments were written in the Latin Script with different types of code-mixing. The language tag of the comment were given. The challenge of the task was to identify the language of the given comment. It was

[2] http://urn.fi/urn:nbn:fi:lb-2020102201

Team	DLI	RDI	SMG	ULI	System Description Papers
HeLju			✓		(Scherrer and Ljubešić, 2021)
HWR	✓				(Jauhiainen et al., 2021b)
LAST	✓		✓		(Bestgen, 2021)
NAYEL	✓				
NRC				✓	(Bernier-Colborne et al., 2021)
Phlyers	✓	✓	✓		(Ceolin, 2021)
SUKI		✓			(Jauhiainen et al., 2021a)
UnibucKernel			✓		(Găman et al., 2021)
UPB		✓			(Zaharia et al., 2021)

Table 1: The teams that participated in the VarDial Evaluation Campaign 2021.

a challenging task, since Tamil, Malayalam and Kannada are closely related languages, some of the words being common in all these languages. The participants had to train a system to identify the language of each comment. Our dataset size is 16,672 comments for training and 4,588 for testing. There were three language tags such as Tamil, Malayalam and Kannada. A new category **Not in intended language** was added to include comments written in a language other than the Dravidian languages.

A sample comment from our dataset provided is displayed below. The original sentence was annotated in Tamil and it contains the English word *movie*. The corresponding English gloss is *You will see what is the movie*.

(1) Paka thana poro **movie** la Enna irukunu baki ellam.

4.2 Participants and Approaches

Due to the short time between the announcement of the shared task and the submission deadline, the participation was lower than we expected. Four teams submitted results to the shared task.

Bestgen (2021) proposed a logistic regression model based on *n*-grams of characters with maximum length as features to classify the comments. The authors achieved a high score with simple techniques. The authors also analyzed the results in detail. For more information, the reader should look at the working notes of the author.

Jauhiainen et al. (2021b) submitted results using two models, a Naïve Bayes (NB) classifier with adaptive language models, which was shown to obtain competitive performance in many language and dialect identification tasks, and a transformer-based model, which is widely regarded as the state-of-the-art in a number of NLP tasks. Their first submission was sent in the closed submission track, using only the training set provided by the shared task organisers. In contrast, the second submission is considered to be open, as it used a pre-trained model trained with external data. Their team attained a shared second position in the shared task with the submission based on Naïve Bayes.

4.3 Results

Results for the DLI task are presented in Table 2.

Rank	Team	Run	Macro-F1
1	LAST	1	0.93
	LAST	2	0.92
	LAST	3	0.92
2	HWR	1	0.92
2	NYAEL	1	0.92
	NAYEL	2	0.91
4	Phlyers	1	0.89
	Phlyers	2	0.89
	HWR	2	0.89
	NAYEL	3	0.84

Table 2: The results of all entries by the four team participating in the DLI shared task in terms of Macro-F1.

Given the difficulty of the DLI 2021 task, the level of performance achieved by the systems is appreciable. Identifying the Other-language category was particularly difficult because it may be thought that it is not homogeneous but composed of different languages in varying proportions. It is not even certain that all the other languages present in the test set were also present in the learning set. Logistic Regression and Naive Bayes methods were used to win the competition. Regarding the systems proposed by Jauhiainen et al. (2021b), even though the difference in performance between

the NB model and the transformers was only 3% on the test set, the fact that the transformers did not outperform the simple NB classifier deserves special attention. One of the reasons behind the inferior performance of the pre-trained models is that the comments contain code-mixed sentences, which were not seen before by pre-trained language models such as BERT or XLM-R.

4.4 Summary

We are glad to see non-native speakers of Dravidian language participating in the DLI task. The DLI shared task showed the difficulty of identifying language in a code-mixed setting. We will continue to add more data to the DLI dataset to improve the language identification for the Dravidian languages in the code-mixed settings in the future.

5 Romanian Dialect Identification (RDI)

5.1 Dataset

As training data, we used an extended version of the Moldavian and Romanian Dialectal Corpus (MOROCO)[3] (Butnaru and Ionescu, 2019), which comprises news articles collected from the top five news websites from Romania and the Republic of Moldova. To automatically annotate the news articles with dialect labels, Butnaru and Ionescu (2019) used the web domains (*.md* or *.ro*) of the news websites. As development data, we used the short text samples from MOROCO-Tweets[4] (Găman and Ionescu, 2020b). The tweets were collected from Romania and the Republic of Moldova, the labels being assigned according to the geographical location. As test data, we collected a new set of tweets, which was compiled in the same manner as MOROCO-Tweets. With these choices as training, development and test corpora, we can evaluate participants on a challenging cross-genre binary dialect identification task: Moldavian (MD) vs. Romanian (RO). The number of samples in the training, the development and the test sets are shown in Table 3. All text samples were automatically pre-processed to replace each named entity with the special token NE.

Dialect	Training	Development	Test
Moldavian	18,121	2,612	2,665
Romanian	21,366	2,625	2,617
Total	**39,487**	**5,237**	**5,282**

Table 3: Number of text samples in the training, the development and the test sets of the RDI shared task.

5.2 Participants and Approaches

Phlyers: The Phlyers (Ceolin, 2021) submitted two runs based on a simple convolutional neural network (CNN). The CNN is first trained on news articles from the official training set, and then fine-tuned on tweets from the official development set. For the first run, the team performed data augmentation by creating ten additional versions of the development set, where the words in each sentence are shuffled. For the second run, the model is trained with even more data augmentation. Both submissions are closed.

SUKI: The predictions submitted by the SUKI team (Jauhiainen et al., 2021a) were produced by a custom coded language identifier based on the product of relative frequencies of character *n*-grams. The model is essentially a Naïve Bayes classifier that uses the relative frequencies as probabilities (Jauhiainen et al., 2019c). The length of the character *n*-grams ranges from 2 to 5. SUKI summed up the negative logarithms of the relative frequencies instead of multiplying them. As a smoothing value, they used the negative logarithm of an *n*-gram appearing only once multiplied by a penalty modifier equal to 1.61. SUKI submitted two closed runs in which they used 50% of the development data as training material and the other 50% for hyperparameter tuning. For the first run, in addition to the basic classifier, they used a blacklist of lowercase character *n*-grams generated from the training and the development data. For the second run, they added the language model adaptation technique described by (Jauhiainen et al., 2018). They used one epoch of language model adaptation to the test data.

UPB: The UPB team (Zaharia et al., 2021) submitted three open runs. For the first run, UPB fine-tuned a Romanian BERT model on the training set, which was split into sentences. After the initial training, they filtered the training set considering only the entries that the model correctly predicted with high confidence for the second round of training. At the same time, they used a prediction threshold to classify an entry as Moldavian or

[3] https://github.com/butnaruandrei/MOROCO

[4] https://github.com/raduionescu/MOROCO-Tweets

Romanian. For the second run, UPB proposed an ensemble formed by training or fine-tuning multiple models, including a Romanian BERT based on adversarial training, a distilled model as well as a method based on generative adversarial networks. For the third run, they submitted the predictions of a student model resulting from knowledge distillation using TextBrewer on Romanian BERT.

5.3 Results

Rank	Team	Run	Macro-F1
1	SUKI	2	0.777182
2	UPB	2	0.732467
	UPB	1	0.731909
	SUKI	1	0.726556
	UPB	3	0.674343
3	Phlyers	1	0.653171
	Phlyers	2	0.513287

Table 4: Macro F_1 scores attained by the teams participating in the 2021 RDI shared task.

As shown in Table 4, the best results in the 2021 RDI shared task were attained by the SUKI team. Compared with their own results Jauhiainen et al. (2020a) obtained in the first edition of the RDI shared task (Găman et al., 2020), the SUKI team improved their performance by a considerable margin. It seems that the main drivers for improvement were (i) the decision to use the development data for training and (ii) the idea of adapting the language model to the test set. The team that was ranked in the second place is UPB. Their best submission is an ensemble that comprises several deep models, including a Romanian BERT. Different from their last year's participation (Zaharia et al., 2020), they carefully split the training set into sentences. This idea was borrowed from top-ranked teams of the 2020 RDI shared task. Phlyers ranked on the third place in the 2021 ranking, without significant differences in terms of performance with respect to their previous participation (Ceolin and Zhang, 2020). Despite having access to significantly more in-domain data compared with the previous RDI shared task, the participants were not able to report significant performance gains. Indeed, the top scoring team (Çöltekin, 2020) in 2020 reached a macro F_1 score of 0.7876, while the top scoring team in 2021 achieved a macro F_1 score of 0.7772. Although the test sets are not identical, we expect them to be equally difficult, since they were collected in the same manner. We thus conclude that Romanian dialect identification remains a difficult task when it comes to short text samples such as tweets, even when in-domain data is available.

5.4 Summary

For the Romanian Dialect Identification shared task, we proposed a cross-domain binary classification task. We had a total of 8 submissions coming from 3 different teams. Each team submitted between 2 and 3 runs. Compared with the 2020 RDI shared task, we observed a decreased interest which can be attributed to the extremely short time given to participants for model development. Looking at the results, we conclude that the set of 5 thousand in-domain text samples (MOROCO-Tweets) can compensate for the much larger set of out-of-domain training samples (MOROCO). However, we did not observe any significant performance boosts compared with last year's RDI shared task, in which the in-domain data available for development was scarce.

6 Social Media Variety Geolocation (SMG)

6.1 Dataset

The SMG task is based on three datasets from two Social Media platforms, Jodel and Twitter. Since its first edition in 2020, the datasets have been expanded.

- The **BCMS subtask** is focused on geolocated tweets published in the area of Croatia, Bosnia and Herzegovina, Montenegro and Serbia in the Serbo-Croatian macrolanguage (ISO acronym HBS, code 639-3). While the training and development data comes from the pool of the 2020 data (Ljubešić et al., 2016), new data collected during 2020 is used for the test set. The training and development data is also divided by the time of publication, the whole train:dev:test setup thereby being significantly more realistic in this iteration of the subtask.

- The **DE-AT subtask** focuses on Jodel conversations initiated in Germany and Austria, which are written in standard German but commonly contain regional and dialectal forms. The training, development and test sets are

created by resampling the 2020 dataset (Hovy and Purschke, 2018).[5]

- The **CH subtask** focuses on Jodel conversations from Switzerland, which were found to be held majoritarily in Swiss German dialects. This dataset is considerably smaller, but we expect it to contain more dialect-specific cues than the DE-AT one. The training, development and test sets are created by resampling the 2020 dataset (Hovy and Purschke, 2018).

All three subtasks use the same data format: each instance consists of three fields, the unprocessed text of the message or conversation, the latitude coordinate and the longitude coordinate. Table 5 shows the key figures of the datasets.

Subtask	Number of instances			Tokens /
	Training	Devel.	Test	instance
BCMS	353,953	38,013	4,189	13
DE-AT	318,487	29,122	31,515	69
CH	25,261	2,416	2,438	50

Table 5: SMG datasets.

6.2 Participants and Approaches

Unfortunately, the participation was much lower than in 2020, due to the short time between the announcement of the shared task and the submission deadline: one team (HeLju) submitted to all three subtasks, whereas another team (UnibucKernel) submitted only to the CH subtask.

HeLju: The HeLju systems (Scherrer and Ljubešić, 2021) rely on the BERT architecture, where the classification output is replaced by a double regression output. HeLju proposes constrained submissions, for which the BERT models are trained from scratch using the SMG training data, as well as unconstrained submissions, for which pre-trained models are used.

UnibucKernel: The UnibucKernel team (Găman et al., 2021) submitted an ensemble system based on XGBoost, whose components are a ν-SVR model trained on top of n-gram string kernels, a CNN with character-level and word-level filters, and a pre-trained BERT model. All components provide double regression outputs. The model represents an upgrade of the previously proposed ensemble (Găman and Ionescu, 2020a) for the 2020 SMG-CH geolocation shared task.

[5]Unfortunately, the Jodel API does not currently allow the collection of new data.

6.3 Results

The test set predictions were evaluated on the basis of median and mean distance to the gold coordinates. Submissions are ranked by decreasing median distance, which is the official metric. For comparison, we also mention the distance values obtained from a simple centroid baseline, which predicts the center point (measured on the training data) for each test instance. Results and rankings for the three tasks are presented in Table 6.

Subtask	Rank	Submission	Median dist. (km)	Mean dist. (km)
BCMS	1	HeLju unconstr.	15.49	76.04
	2	HeLju constr.	52.06	98.74
		Baseline	*118.33*	*160.78*
DE-AT	1	HeLju unconstr.	149.33	172.52
	2	HeLju constr.	161.13	184.97
		Baseline	*206.42*	*226.13*
CH	1	HeLju unconstr.	17.55	25.84
	2	HeLju constr.	20.70	29.62
	3	UnibucKernel	23.60	29.75
		Baseline	*53.13*	*51.50*

Table 6: SMG task results. The official metric is median distance in kilometers, i.e., lower values are better.

The low number of submissions does not allow us to draw reliable conclusions, but the general findings are similar to last year's: the CH subtask turned out to be the easiest one and the DE-AT the most difficult one, with BCMS lying between the two. All submissions managed to beat the baseline by a large margin, and unconstrained systems again tend to beat constrained ones.

The median distance value of the best-ranked BCMS submission seems surprisingly low. The reason for this outlier is probably to be found in the way the 2021 data were obtained. The test set consists entirely of tweets published after March 2020. Thus, it is likely that the limitation in population movements due to COVID restrictions led to a more skewed geographical distribution of the test instances, which in turn makes it easier to reach low median values.

6.4 Summary

The second edition of the SMG task attracted fewer participants than the first, and as a consequence, the variety of explored solutions and algorithms is also narrower. Nevertheless, we believe that a geolocation task has its justification within VarDial, in particular for pluricentric languages without clear-cut variety borders. Thanks to its reliance on easily available geolocated messages from social media services, future editions of the SMG task can be envisaged, possibly focusing on different language areas.

7 Uralic Language Identification (ULI)

The ULI shared task focuses on 29 rare Uralic languages and especially the difficulties of finding such languages amongst a huge amount of textual material in more common languages. In ULI, the 29 rare Uralic languages are considered relevant. In addition to them, there are 149 non-relevant languages.

The shared task was first organized as part of the VarDial Evaluation Campaign 2020 (Găman et al., 2020). Due to low participation, we decided to keep the shared task open even after the campaign was over. Only the NRC team had submitted results and they were all well below the baseline. We constructed a leaderboard page with the best results updated as soon as they were evaluated.

The ULI 2021 shared task contained three separate subtasks: **ULI-RLE**, **ULI-RSS**, and **ULI-178**. In **ULI-RLE** (relevant languages as equals), the defining measure was the macro F_1 score calculated for the relevant languages present in the training set. In **ULI-RSS** (relevant sentences as equals), the measure used was the micro F_1 score calculated for sentences either written in or predicted to be written in the relevant languages. In **ULI-178** (All 178 languages as equals), the macro F_1 score was calculated as an average over all the 178 languages part of the training set repertoire.

We were accepting submissions until the end of the evaluation phase of the VarDial Evaluation Campaign 2021 on February 2, 2021. Participants who submitted results were all invited to submit a system description paper to appear in the proceedings of VarDial 2021.

7.1 Dataset

The dataset for the 2021 competition was the same as earlier. It is described by Găman et al. (2020) and in more detail by Jauhiainen et al. (2020b). In short, the training set consists of two parts, the relevant and the non-relevant languages. The training data for the relevant languages comes from the Wanca 2016 collection (Jauhiainen et al., 2019a).[6] Wanca 2017 containing the test data for relevant languages remains unpublished, but the publication is expected to occur in 2021. The training and the test data for the non-relevant languages comes from the Leipzig Corpora Collection (Richter et al., 2006).[7] The ULI leaderboard contains links to the download locations of the training and the test sets.

7.2 Participants and Approaches

Three teams submitted new results during the ULI 2021 evaluation period.

NRC: The NRC team was the only one submitting results for the initial ULI shared task (Bernier-Colborne and Goutte, 2020). In 2020, they used a system based on the one they had used to win the Cuneiform Language Identification (CLI) shared task in the 2019 VarDial Evaluation Campaign (Jauhiainen et al., 2019b; Bernier-Colborne et al., 2019; Zampieri et al., 2019). However, the ULI task turned out to be much more difficult for the BERT based language classifier system. For the ULI 2021, they set out to further improve the results produced by the deep learning architecture (Bernier-Colborne et al., 2021). In addition, they submitted results using a probabilistic classifier similar to Naïve Bayes. They used this NB style classifier already in the DSL shared task of 2014 (Zampieri et al., 2014) to predict the language group before handing the task over to SVMs (Goutte et al., 2014).

LAST: The LAST team submitted several runs using Logistic Regression (LR) classifiers and their ensembles (Bestgen, 2021). As features, the classifiers used character *n*-grams, either word internal or word spanning, which were weighted with BM25. BM25 weighted character *n*-grams were used successfully by the CECL team in Discriminating between Similar Languages (DSL) and GDI shared tasks of VarDial 2017 (Bestgen, 2017). Then they were used as features for their SVM based classifiers.

[6] http://urn.fi/urn:nbn:fi:lb-2020022901

[7] https://corpora.uni-leipzig.de/

Rank	Team	Method	Relevant Macro-F1
1	NRC	Probabilistic classifier (similar to Naive Bayes) using character 5-grams	0.8138
2	Phlyers	Ensemble of SVM and Naive Bayes classifiers using character *n*-grams 3-5.	0.8085
	baseline	HeLI	0.8004
	Phlyers	Naive Bayes classifier trained on character 5grams	0.7977
3	LAST	Ensemble of LR classifiers trained on char *n*-grams 1-3 weighted with BM25	0.7758
	LAST	LR classifier trained on char *n*-grams 1-3 weighted with BM25	0.7755
	Phlyers	SVM (char *n*-grams 3-4) followed by Naive Bayes classifier (char *n*-grams 3-5)	0.7740
	LAST	LR classifier trained on word internal char *n*-grams 1-4 weighted with BM25	0.7727
	Phlyers	Naive Bayes classifier trained on character 3grams and 4grams	0.7584
	NRC	BERT-style deep neural network with early stopping	0.7430
	NRC	BERT-style deep neural network	0.6866
	Phlyers	SVM (char *n*-grams 5-7) followed by Naive Bayes classifier (char *n*-grams 3-5)	0.6783

Table 7: ULI shared task 2021 - RLE results.

Rank	Team	Method	Relevant Micro-F1
1	NRC	Probabilistic classifier (similar to Naive Bayes) using character 5-grams	0.9668
	baseline	HeLI	0.9632
	NRC	BERT-style deep neural network with early stopping	0.9530
2	LAST	Ensemble of LR classifiers trained on char *n*-grams 1-3 weighted with BM25	0.9496
	LAST	LR classifier trained on word internal char *n*-grams 1-4 weighted with BM25	0.9492
	LAST	LR classifier trained on char *n*-grams 1-3 weighted with BM25	0.9484
3	Phlyers	SVM (char *n*-grams 3-4) followed by Naive Bayes classifier (char *n*-grams 3-5)	0.8389
	NRC	BERT-style deep neural network	0.8177
	Phlyers	SVM (char *n*-grams 5-7) followed by Naive Bayes classifier (char *n*-grams 3-5)	0.7595
	Phlyers	Naive Bayes classifier trained on character 5grams	0.5934
	Phlyers	Ensemble of SVM and Naive Bayes classifiers using character *n*-grams 3-5.	0.5932

Table 8: ULI shared task 2021 - RSS results.

Rank	Team	Method	Macro-F1
	baseline	HeLI	0.9252
1	LAST	LR classifier trained on word internal char *n*-grams 1-4 weighted with BM25	0.9164
	LAST	Ensemble of LR classifiers trained on char *n*-grams 1-3 weighted with BM25	0.9131
	LAST	LR classifier trained on char *n*-grams 1-3 weighted with BM25	0.9125
2	NRC	Probabilistic classifier (similar to Naive Bayes) using character 5-grams	0.9079
	NRC	BERT-style deep neural network with early stopping	0.9039
3	Phlyers	Ensemble of SVM and Naive Bayes classifiers using character *n*-grams 3-5.	0.8847
	Phlyers	Naive Bayes classifier trained on character 5grams	0.8831
	Phlyers	Naive Bayes classifier trained on character 3grams and 4grams	0.8753
	NRC	BERT-style deep neural network	0.8366

Table 9: ULI shared task 2021 - 178 results.

Phlyers: The Phlyers team used different combinations and ensembles of Naïve Bayes and SVM classifiers (Ceolin, 2021). As features, they used varying sized character *n*-grams. They experimented with similar systems in their submissions to the RDI shared task in 2020 (Ceolin and Zhang, 2020).

7.3 Results

Tables 7, 8, and 9 show the VarDial 2021 Evaluation Campaign end results for the ULI 2021 competition. As baseline, we used a HeLI based language identifier using parameters presented by Jauhiainen et al. (2017). Relative frequencies were calculated from the training data for character *n*-grams and words, but the baseline was not tuned using the training set.

In the ULI-RLE subtask, both the NRC and the Phlyers teams managed to beat the HeLI baseline. The NRC team's probabilistic classifier using character 5-grams is the new state of the art in ULI-RLE. In ULI-RSS, the Phlyers teams submission was not competetive at all, but the NRC team surpassed the baseline with the same system as they had used to win the ULI-RLE. For the third subtask, ULI-178, the submitted results failed to improve on the strong

baseline provided by the HeLI based language identifier. The best submission was by the LAST team using an LR classifier with BM25-weighted word internal character *n*-grams.

7.4 Summary

We were glad to see more active participation in the ULI task than during the previous Evaluation Campaign. The ULI shared task proved again to be too difficult for the deep learning based classifiers and the more traditional approaches won all the subtasks. We will continue to keep the ULI leaderboard open and the results list can be updated again after the VarDial 2021 workshop is over. During 2021, we are aiming to produce a joint error analysis of several systems which have participated so far and design a new dataset for ULI 2022.

8 Conclusion

This paper presented the results and findings of the four shared tasks organized as part of the VarDial Evaluation Campaign 2021: Dravidian Language Identification (DLI), Romanian Dialect Identification (RDI), Social Media Variety Geolocation (SMG), and Uralic Language Identification (ULI). Each of these tasks addressed an important challenge in language and dialect identification providing participants with either new or augmented versions of existing datasets that are made freely available to the community.

We included short descriptions for each team's systems in this report and references to all 8 system description papers published in the VarDial workshop proceedings in Table 1. Despite the state-of-the-art performance obtained by deep learning models in a wide range of NLP tasks, in VarDial 2021 we observed that traditional machine learning models once again outperformed deep learning models for language and dialect identification. This corroborates the findings of previous editions of the campaign (Zampieri et al., 2019; Găman et al., 2020) and of the survey by Jauhiainen et al. (2019d).

Acknowledgments

We thank all the participants for their interest in the evaluation campaign.

The work related to the ULI shared task has been partly funded by the Kone Foundation, The Finnish Research Impact Foundation, and the University of Helsinki in cooperation with Lingsoft.

References

Gabriel Bernier-Colborne and Cyril Goutte. 2020. Challenges in neural language identification: NRC at VarDial 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 273–282, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

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.

Gabriel Bernier-Colborne, Serge Léger, and Cyril Goutte. 2021. N-gram and neural models for uralic language identification: NRC at VarDial 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Yves Bestgen. 2017. Improving the character ngram model for the dsl task with bm25 weighting and less frequently used feature sets. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 115–123, Valencia, Spain.

Yves Bestgen. 2021. Optimizing a Supervised Classifier for a Difficult Language Identification Problem. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Andrei M. Butnaru and Radu Tudor Ionescu. 2019. MOROCO: The Moldavian and Romanian Dialectal Corpus. In *Proceedings of ACL*, pages 688–698.

Andrea Ceolin. 2021. Comparing the performance of CNNs and shallow models for language identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Andrea Ceolin and Hong Zhang. 2020. Discriminating between standard Romanian and Moldavian tweets using filtered character ngrams. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 265–272, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Bharathi Raja Chakravarthi. 2020. *Leveraging orthographic information to improve machine translation of under-resourced languages*. Ph.D. thesis, NUI Galway.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and*

Collaboration and Computing for Under-Resourced Languages (CCURL), pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Çağrı Çöltekin. 2020. Dialect Identification under Domain Shift: Experiments with Discriminating Romanian and Moldavian. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 186–192, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The nrc system for discriminating similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 139–145, Dublin, Ireland.

Mihaela Găman, Sebastian Cojocariu, and Radu Tudor Ionescu. 2021. UnibucKernel: Geolocating Swiss German Jodels Using Ensemble Learning. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Mihaela Găman, 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 Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Mihaela Găman and Radu Tudor Ionescu. 2020a. Combining Deep Learning and String Kernels for the Localization of Swiss German Tweets. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 242–253.

Mihaela Găman and Radu Tudor Ionescu. 2020b. The Unreasonable Effectiveness of Machine Learning in Moldavian versus Romanian Dialect Identification. *arXiv preprint arXiv:2007.15700*.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.

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.

Heidi Jauhiainen, Tommi Jauhiainen, and Krister Linden. 2019a. Wanca in Korp: Text corpora for underresourced Uralic languages. In *Proceedings of the Research data and humanities (RDHUM) 2019 conference*, number 17 in Studia Humaniora Ouluensia, pages 21–40, Finland. University of Oulu.

Tommi Jauhiainen, Heidi Jauhiainen, Tero Alstola, and Krister Lindén. 2019b. Language and dialect identification of cuneiform texts. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 89–98. Association for Computational Linguistics.

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. 2020a. Experiments in Language Variety Geolocation and Dialect Identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 220–231, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2021a. Naive Bayes-based Experiments in Romanian Dialect Identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen, and Krister Lindén. 2020b. Uralic Language Identification (ULI) 2020 shared task dataset and the Wanca 2017 corpus. In *Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 688–698.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2017. Evaluation of Language Identification Methods Using 285 Languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa 2017)*, pages 183–191, Gothenburg, Sweden. Linköping University Electronic Press.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019c. 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. 2019d. Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Tommi Jauhiainen, Tharindu Ranasinghe, and Marcos Zampieri. 2021b. Comparing Approaches to Dravidian Language Identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A Survey of Current Datasets for Code-Switching Research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

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.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P. McCrae. 2020. Named Entity Recognition for Code-Mixed Indian Corpus using Meta Embedding. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 68–72.

Matthias Richter, Uwe Quasthoff, Erla Hallsteinsdóttir, and Christian Biemann. 2006. Exploiting the leipzig corpora collection. In *Proceedings of the Information Society Language Technologies Conference*, Ljubljana.

Yves Scherrer and Nikola Ljubešić. 2021. Social media variety geolocation with geobert. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

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).

George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. 2021. Dialect Identification through Adversarial Learning and Knowledge Distillation on Romanian BERT. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. 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)*, Valencia, Spain.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shuon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Santa Fe, USA.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei Butnaru, and Tommi Jauhiainen. 2019. A Report on the Third VarDial Evaluation Campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the dsl shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 58–67, Dublin, Ireland.