Early Guessing for Dialect Identification

Vani Kanjirangat
IDSIA-USI/SUPSI, Switzerland
vanik@idsia.ch

Tanja Samardzic
URPP Language and Space, UZH
tanja.samardzic@uzh.ch

Fabio Rinaldi
IDSIA-USI/SUPSI, Switzerland
fabio.rinaldi@idsia.ch

Ljiljana Dolamic
armasuisse S+T, Switzerland
Ljiljana.Dolamic@armasuisse.ch

Abstract

This paper deals with the problem of incremental dialect identification. Our goal is to reliably determine the dialect before the full utterance is given as input. The major part of the previous research on dialect identification has been model-centric, focusing on performance. We address a new question: How much input is needed to identify a dialect? Our approach is a data-centric analysis that results in general criteria for finding the shortest input needed to make a plausible guess. Working with three sets of language dialects (Swiss German, Indo-Aryan and Arabic languages), we show that it is possible to generalize across dialects and datasets with two input shortening criteria: model confidence and minimal input length (adjusted for the input type). The source code for experimental analysis can be found at Github [1].

1 Introduction

Language identification depends very much on what kind of languages we are discriminating. If languages to be discriminated are distant (e.g., Russian vs. Chinese), the task is straightforward and a short sequence of words provides enough information to assign the correct class. But if languages are similar and written in the same script (e.g. Russian vs. Ukrainian), much longer samples are needed to encounter the discriminating features (Tiedemann and Ljubešić, 2012). The task is even harder when dealing with non-standard orthography, which we find in written dialects and user posts on the internet (Zampieri et al., 2017).

Current research is mostly concerned with improving the performance of the task by applying increasingly sophisticated methods, including pretrained models, whose performance varies across different language dialects (Jauhiainen et al., 2021). However, many other aspects of the task may play an important role in practical applications. One of such challenges is the possibility to make early guesses on the language or dialect before seeing the whole message. Such a feature can be especially useful for more dynamic classification of a continuous stream of messages as in transcribed speech or instant messaging, where sentence boundaries are not clearly defined and the input can be segmented arbitrarily. A reliable early classification can improve further processing (e.g., translation, information extraction).

In this paper, we address the problem of early guessing in dialect identification mostly from the data-centric point of view, but also consider some model-centric issues. We search for criteria for shortening the input so that the model performance is the same or similar to the performance obtained with the full input. We perform experimental studies with four datasets representing three sets of dialects with considerably different writing practices (some writings are more standard than others) and find that the same input shortening criteria give the best results in all the cases.

2 Related Work

The task of dialect identification and discrimination between similar languages is mostly addressed in the scope of the VarDial Evaluation Campaign (Zampieri et al., 2017, 2018, 2019). The organizers of the tasks released datasets for various dialects and similar languages, such as Swiss-German, Indo-Aryan, Uralic, Romanian, Arabic, Chinese, etc. In Arabic dialect identifications (ADI), other popular datasets include Arabic Online Commentary (AOC) (Zaidan and Callison-Burch, 2011), Nuanced Arabic Dialect Identification (NADI) Twitter data set (Abdul-Mageed et al., 2020)[2], Multi-Arabic Dialect Applications and Re-

[1] https://github.com/vanikanjirangat/Dialect_Early_Guessing

[2] https://sites.google.com/view/nadi-shared-task

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6417–6426
December 7-11, 2022

sources (MADAR)(Bouamor et al., 2019)[3] etc.

With the advent of transformers-based models, we see a wide use of pre-trained models in dialect classifications (Popa and Ștefănescu, 2020; Zaharia et al., 2020; Ljubešić and Lauc, 2021), but traditional approaches based on n-gram statistics also seems to remain quite popular (Ali, 2018; Ciobanu et al., 2018b; Jauhiainen et al., 2018; Gupta et al., 2018; Çöltekin et al., 2018; Ciobanu et al., 2018a; Bernier-Colborne et al., 2021). In the standard datasets (data from written sources) of the VarDial task, such as Indo-Aryan and Romanian, language specific pre-trained models worked well (Zaharia et al., 2020, 2021), while in the non-standard datasets (transcribed from speeches), such as Swiss German and Arabic, pre-trained transformer models didn't boost the performance as expected. Regarding other Arabic datasets, viz., AOC and NADI, neural (CNN, BiLSTM, etc.) and transformer models were successfully used (Elaraby and Abdul-Mageed, 2018; Talafha et al., 2020; Zhang and Abdul-Mageed, 2019; Beltagy et al., 2020).

In contrast to most previous work, the main focus is not on improving the performance when the task is the classification of the whole utterance but on finding the minimal input on which an acceptable classification performance can be achieved. This aspect of the problem has been so far only minimally addressed in the literature. One such study was on the influence of the length of the utterances in ADI on Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007) for coarse and fine-grained dialect classification using n-gram features with Multinomial Naive Bayes classifier. They found that with an average length of 7 words, it is possible to classify the dialects with acceptable accuracy (Salameh et al., 2018). Our study is a more general data-centric exploration of the early classification involving varied datasets in multiple languages.

3 Data

For our study, we select four datasets, three offered by the VarDial Evaluation Campaign (see Section 2). We perform three tasks: German Dialect Identification (GDI)[4], Indo-Aryan Language Identification (ILI)[5] and the Arabic Dialect Identification.[6]

	GDI	**ILI**	**ADI VarDial**	**ADI AOC**
Train	14647	68453	21001	86541
Dev	4659	8286	1566	10820
Test	4752	9032	1492	10812

Table 1: The size of datasets expressed as the number of utterances.

The GDI dataset represents four areas of Swiss German: Basel, Bern, Lucerne, and Zurich. Training and the test datasets are obtained from the ArchiMob corpus of Spoken Swiss German (Samardzic et al., 2016). GDI datasets are available from the years 2017-2019. We work with the GDI-2018 in the 4-way classification setting. The ILI task identifies five closely related languages from the Indo-Aryan language family, namely, Hindi, Braj Bhasha, Awadhi, Bhojpuri, and Magahi. For each language, 15,000 sentences are extracted mainly from the literature domain. The sources were previously published either on the internet or in print. These languages are often mistakenly considered to be varieties of Hindi. The ADI VarDial task (Malmasi et al., 2016; Ali et al., 2016) focused on five classes, viz., Modern Standard Arabic (MSA), Egyptian (EGY), Gulf (GLF), Levantine (LAV), Moroccan (MOR), and North-African (NOR). MSA is the modern variety of the language used in news and educational articles. This differs from the actual communication language of native speakers lexically, syntactically and phonetically. The VarDial ADI dataset is both speech transcribed and transliterated to English. Another Arabic dialect dataset used for the experimentation is the Arabic Online Commentary (AOC) (Zaidan and Callison-Burch, 2011) dataset. This constitutes a large-scale repository of Arabic dialects and covers MSA and the dialectal varieties, viz., Egyptian (EGY), Gulf (GLF), Levantine (LEV), and Moroccan (MOR). Table 1 reports the data statistics of GDI, ILI and ADI datasets used for our experimentation. We represent the ADI dataset from VarDial as ADI-VarDial and AOC as ADI-AOC, respectively.

[3] https://camel.abudhabi.nyu.edu/madar-shared-task-2019/

[4] https://drive.switch.ch/index.php/s/DZycFA9DPC8FgD9

[5] https://github.com/kmi-linguistics/VarDial2018

[6] https://arabicspeech.org/resources/

4 Methods

We perform incremental analysis by running the same classifier on varied substrings of the test input. This is to understand the performance at different lengths of the input text. We start with the first word, then repeat the classification with the first two words and so on until we reach the end of the utterances. We refer to all the incremental substrings as fragments. We observe the model's performance at each incremental step and analyze its state (confidence) to determine the earliest point when a plausible guess can be made. We perform extensive analysis on the influence of different factors that directly and indirectly affect the model performance after applying the shortening criteria.

Models In the case of each dialect group, we compare the base model (pretrained on English) with one or more models pretrained on a language more closely related to the target dialect group. For the GDI data set, we compared three models: BERT-base-cased model (Devlin et al., 2019), multilingual BERT (mBERT) and German BERT[7]. In the case of the ILI dataset, we compared four models: BERT-base-cased, mBERT, IndicTransformers (Jain et al., 2020) and IndicBERT(Kunchukuttan et al., 2020). IndicBERT covers 12 languages, including Hindi, Tamil, English, Malayalam, etc., trained using AI4Bharat's corpus and is based on multilingual ALBERT.[8] IndicTransformers[9] is a BERT model trained with 3 GB of data from the OSCAR corpus [10] and covers Hindi, Bengali and Telugu. For ADI-VarDial, we used two main models: BERT-base-cased and AraBERT (Antoun et al., 2020).[11]. We did the experiments with mBERT too, which gave poor performance (accuracy of only 28% with 7% F-score) and hence did not carry out further experiments with mBERT in ADI-Vardial. AraBERT is based on BERT-base model; it is additionally pre-trained on Arabic news articles and two publicly available large Arabic corpora covering 24 Arab countries. For the ADI-AOC dataset, we compare three models: BERT-base-cased, mBERT and AraBERT.

Input Shortening We first tokenize the input sentence by splitting it into white spaces. We then create fragments that consist of incrementally increased prefixes of the original utterance. The length of fragments ranges between 1 and N, where N is the original utterance's length (in tokens). For example, consider the test sentence: *'das haisst im klarteggst'* of length N=4. The incremental fragments will be: [*'das'*, *'das haisst'*,*'das haisst im'*,*'das haisst im klarteggst'*].
The number of fragments obtained in each case is listed in Appendix A. For each fragment, we obtain predictions using the same fine-tuned model. We collect the information about model prediction and its confidence for further analyses.

Model Confidence Analysis with Temperature Scaling The confidence scores of a model can be very high (close to 1) even when the predictions are incorrect. Calibration is a method to disincentivize a model from being over-confident (Bella et al., 2010; Nixon et al., 2019; Widmann et al., 2019). Although the transformers models are considered to be well-calibrated (Desai and Durrett, 2020), methods such as temperature scaling (Guo et al., 2017) and label smoothing (Müller et al., 2019) can improve the calibration. We expect this to help, especially for the case of GDI data, where the overall performance is rather low compared to the other datasets. We explore temperature scaling to calibrate the prediction probabilities of our model: we divide the non-normalized logits (before the softmax operation) with the scalar temperature hyperparameter T. After this step, the prediction probability is obtained using the usual Softmax function. In exploring model confidence as a shortening criterion, we use calibrated probabilities. The details of the model calibration are explained in Appendix C.

5 Experiments and Results

Each model was trained for 4 epochs with Adam optimizer using a learning rate of 2e-5 on the corresponding training set using 1 Tesla K80 GPU. We used the pre-trained models from the HuggingFace library.[12]

BERT-base vs. Linguistic Proximity Table 2 shows the classification accuracy with full input and with shortened input based on the input shortening criteria (Explained in Section 5 and Appendix D). We note that the ILI and ADI-AOC datasets are the closest to standard writing, while much

[7] https://www.deepset.ai/german-bert
[8] https://indicnlp.ai4bharat.org/indic-bert/
[9] https://huggingface.co/neuralspace-reverie
[10] https://oscar-corpus.com/
[11] https://huggingface.co/aubmindlab/bert-base-arabert
[12] https://huggingface.co/models

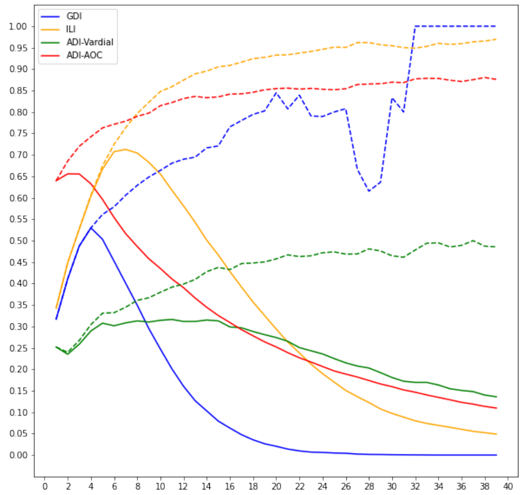

Figure 1: Proportion of Correct Predictions at each Fragment Length (%) for GDI, ILI and ADI Datasets. Solid lines represent the proportion to the (constant) test set data size. Dashed lines show the relative proportion of correct predictions in the set of all test instances of the given length.

Dataset	Model	Full	Short
GDI	BERT-base-cased	**62**	55.2
	mBERT	59	50.8
	German BERT	60	54.4
ILI	BERT-base-cased	81	56.5
	mBERT	88	69.9
	IndicBERT	84	54.4
	IndicTransformers	**90**	73.7
ADI_VarDial	BERT-base-cased	**43**	33.7
	AraBERT	38	29.8
ADI_AOC	BERT-base-cased	81	64.3
	mBERT	82	65.8
	AraBERT	**82**	66.7

Table 2: The accuracy (%) with different pretrained models on full utterances and on shortened input.

more non-standard writing is found in the other two datasets. Linguistic proximity [13] of the pre-trained model seems useful only in combination with standard writing (ILI and ADI-AOC dataset). In ADI-AOC, we can see that AraBERT is more helpful with shortened input than mBERT, although they both achieve similar accuracy with full input. The best performance on the strongly non-standard data (GDI and ADI-VarDial) is still obtained with BERT-base-cased. Note, however, that the best performance on these non-standard datasets is rather low in absolute terms, underlining the need for better models in this domain.

Fragment Length The impact of the fragment lengths on the model is analyzed in Figure 1. The solid lines in the plot represent the ratio of the correct predictions at each fragment length n to the total number of test instances (which is constant for a given corpus). The dashed lines represent the ratio of the correct predictions at fragment length n to the number of possible predictions at that length (i.e. sentences of length n or longer). These graphs show that it is possible to make accurate and useful predictions with shorter fragments, and the additional gain with longer fragments is proportionally lower. A first peak for the GDI data is obtained at the length 4, while the peak is on the length 7 for the ILI dataset. In ADI-AOC, the peak is on the length 2. The trend in all these cases is the same, modulated by the length of the original utterances (longer in ILI). The trends are somewhat different in the ADI-Vardial dataset: there are no clear peaks, but the best scores are still obtained on shorter segments.

Input Shortening Criteria For each set of dialects, we found the minimum input length for best performances by repeated experiments on different fragment lengths. We found that this minimum fragment length was same as the peak values in Figure 1. For ADI-Vardial, we found that at length 8, the maximum performance is obtained. Further, the maximum input length is determined according to two additional criteria: model confidence after temperature scaling and label consistency. We experiment with several criteria defined in terms of these two variables (described in detail in Appendix D).We find that the same selection criterion gives the best results in all the data sets: the first decrease in the model confidence after the minimum length threshold. In addition to this, label consistency was helpful in all the data sets except GDI: the results improve when we consider the decrease in the model confidence only when the current and the previous labels are the same. While analyzing the performance on the shortened input, we checked what would happen in an ideal case if we knew where to cut the input utterance in each case. This performance, which is the upper bound on our task is, in fact, better than the performance with the full input (details in Appendix B). It provides an empirical justification for the research goal of finding criteria for shortening the input.

6 Conclusion

We have identified general criteria for making early guesses in dialect identification: language spe-

[13] We refer to linguistic proximity in context with the language-specific pre-trained models

cific minimal length of the input and language-independent change in the model confidence score (the first decrease in the confidence score). While these criteria do not maintain the performance achieved with the full input, they are the starting point for further optimization, which can eventually lead to an overall improvement on the task.

7 Limitations

The main limitation of our work is the fact that early guesses do not achieve the performance as the full input. We have shown empirically that it is possible to maintain or even improve the performance with early guessing and that this is a goal worth pursuing. Elaborating methods to achieve this goal remains outside of the scope of the current paper and we leave it for future work. Another limitation concerns the generalization of our findings to standard and non-standard data, which still needs to be better understood.

References

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. The shared task on nuanced arabic dialect identification (nadi). In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP2020), Barcelona, Spain*.

Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals. 2016. Automatic dialect detection in arabic broadcast speech.

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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15.

Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. 2010. Calibration of machine learning models. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 128–146. IGI Global.

Ahmad Beltagy, Abdelrahman Abouelenin, and Omar ElSherief. 2020. Arabic dialect identification using bert-based domain adaptation. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 262–267.

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*, pages 128–134.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The madar shared task on arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207.

Alina Maria Ciobanu, Shervin Malmasi, and Liviu P Dinu. 2018a. German dialect identification using classifier ensembles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 288–294.

Alina Maria Ciobanu, Marcos Zampieri, Shervin Malmasi, Santanu Pal, and Liviu P Dinu. 2018b. Discriminating between indo-aryan languages using svm ensembles. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 178–184.

Çağrı Çöltekin, Taraka Rama, and Verena Blaschke. 2018. Tübingen-oslo team at the vardial 2018 evaluation campaign: An analysis of n-gram features in language variety identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 55–65.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302.

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.

Mohamed Elaraby and Muhammad Abdul-Mageed. 2018. Deep models for arabic dialect identification on benchmarked data. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 263–274.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330.

Divyanshu Gupta, Gourav Dhakad, Jayprakash Gupta, and Anil Kumar Singh. 2018. Iit (bhu) system for indo-aryan language identification (ili) at vardial 2018. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 185–190.

Kushal Jain, Adwait Deshpande, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2020. Indic-transformers: An analysis of transformer language models for indian languages. *arXiv preprint arXiv:2011.02323*.

Tommi Jauhiainen, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Comparing approaches to Dravidian language identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 120–127, Kiyv, Ukraine. Association for Computational Linguistics.

Tommi Sakari Jauhiainen, Heidi Annika Jauhiainen, Bo Krister Johan Linden, et al. 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)*. The Association for Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Avik Bhattacharyya, Mitesh M Khapra, Pratyush Kumar, et al. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Nikola Ljubešić and Davor Lauc. 2021. Bertić-the transformer language model for bosnian, croatian, montenegrin and serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42.

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.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in Neural Information Processing Systems*, 32:4694–4703.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2.

Cristian Popa and Vlad Ștefănescu. 2020. Applying multilingual and monolingual transformer-based models for dialect identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 193–201.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344.

Tanja Samardzic, Yves Scherrer, and Elvira Glaser. 2016. Archimob-a corpus of spoken swiss german. In *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 111–118.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India. The COLING 2012 Organizing Committee.

David Widmann, Fredrik Lindsten, and Dave Zachariah. 2019. Calibration tests in multi-class classification: A unifying framework. *Advances in Neural Information Processing Systems*, 32.

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.

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*, pages 113–119.

Omar 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.

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*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, et al. 2018. Language identification and morphosyntactic tagging. the second vardial evaluation campaign.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, et al. 2019. A report on the third vardial evaluation campaign. Association for Computational Linguistics.

Chiyu Zhang and Muhammad Abdul-Mageed. 2019. No army, no navy: Bert semi-supervised learning of arabic dialects. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 279–284.

A Fragment Statistics

The incremental processing gives 42797 fragments for the 4752 test cases in the GDI dataset. In ILI, we obtain 170710 fragments from 9032 test cases. In ADI-VarDial, 59498 fragments from 1492 test cases were obtained, while 205530 fragments from 10812 test cases were obtained in ADI-AOC.

B Upper Bound

To see whether correct predictions are possible before seeing the full utterance, we first find the minimum length fragment at which a correct prediction is made. For instance, consider the incrementally processed example:

[*'das'*
'das haisst'
'das haisst im'
'das haisst im klarteggst']

At length four (the fourth line in the example above), we obtain a true prediction and hence this will be selected as the optimal shortened input for the given utterance since the predicted class is wrong in the previous three fragments (lines 1-3 in the example above). In this case, length 4 is the shortest length at which the correct prediction is obtained (which could be at an earlier length also). We find such fragments for each original test utterance (one fragment per utterance) and then compute the classification accuracy with respect to these optimal input lengths.

Measured in this way, the accuracy scores are higher than the full-input classification. In the case of GDI, we get 80% (compared to 62% on the full input). For ILI, we obtained an upper bound of 94% compared to the 90% accuracy exhibited by the best baseline model. In ADI-VarDial, the accuracy obtained with a similar setup was 74.46% (compared to 43% with full input). Finally, in ADI-AOC, we obtained an upper bound of 90.16% (compared to 82% with full input). Hence, it was observed that if we can find the optimal early guess point for all the language datasets, the performance improves. We consider this accuracy to be our upper bound: this is what could be achieved if we knew where to cut the input utterance in each case.

C Temperature Scaling

The values of the temperature scaling parameter $T > 0$ are the same for all classes and they are optimized with respect to the Negative-Log-Likelihood (NLL) loss on the validation set. To compare the models after and before calibration, we use Expected Calibration Error (ECE) as shown in Equation (1).

$$ECE = \sum_{k=1}^{K} \frac{b_k}{n} |acc(k) - conf(k)| \qquad (1)$$

Calibration is formally expressed as a joint distribution which can be approximated by binning the predictions to K disjoint sets. Each bin will have b_k predictions and n is the number of samples. ECE is defined as the weighted average of the difference between each bin's accuracy and confidence or posterior probability. A perfectly calibrated model has *conf(k) = acc(k)* for each bucket of real-valued predictions.

We found that ECE decreases considerably with calibration using temperature scaling (TS) for all the fine-tuned models.

For the fine-tuned BERT-base-cased model without TS, the ECE was 23.96, while with TS *t*= 2.28, ECE dropped to 6.3 in the GDI dataset. Similar experiments were done on ILI data with the Indic-Transformer model to fine-tune the *T* value. At t=1, we have an ECE of 20.09 for ILI, while after calibrations at t=1.79, ECE dropped to 13.91. In ADI-VarDial using Bert-base-cased, at t=1, we obtained an ECE of 15.15, which dropped on to 7.45 with the fine-tuned temperature value, t=2.32. For AOC, with AraBERT, the ECE was 11.03 at t=1 and after temperature scaling with t=1.55, ECE reduced to 1.09.

D Explored Early Guessing Possibilities

In Tables 3, 4, 5 and 6, *'current'* is the current fragment under consideration. *prob()* is the calibrated probability. We compare the *prob(current)* with *prob(previous)* and *prob(next)*. As discussed, the fragment is the output of incremental processing. The criteria checks will be done for each group of fragments associated with a particular sentence. Another input shortening criterion included is labeling consistency. Here we check the consistency of predicted labels, *predicted label*. Each of these input shortening criteria is evaluated separately and in combination with each other. We consider the fragment that satisfies the input shortening criteria at the first position after a pre-defined length point, say, *m*. The value *m* will be different for each language and needs to be tuned based on performance metrics (accuracy/ F-score). The same input shortening criteria were evaluated for all the datasets while considering different starting lengths *m*. For GDI we found optimal *m=4*, in ILI *m=7* while *m=8* and *m=2* in ADI-VarDial and ADI-AOC respectively. The results for each input shortening criterion are reported in Tables 3, 4, 5 and 6. All the input shortening criteria are evaluated separately and some of the potential input shortening criteria are evaluated in combination.

In ADI-VarDial and ILI, we observed that the best heuristic was p4 & l1, while in GDI, it was p4. In ADI-AOC, the addition of p1 to p4 improves the performance by 0.21 points. The behaviors were the same across the different models. Generalizing, we found that with p4, the performance is good across all the languages, which means the main criterion is *predicted probability(current)>predicted probability(next)*. In other words, we stop the incremental classification once the model probability starts decreasing. In ILI and ADI-VarDial, the addition of l1 criteria *predicted label(current)==predicted label(next)* adds to the performance by 0.39 points and 0.87 points respectively.

Overall, we observed that it is possible to generalize an input shortening criteria across different language dialects. The slight variations could be due to the nature of the data sets and inherent differences in the languages. For instance, the ILI and ADI-AOC datasets comprised standard written texts, while GDI and ADI-VarDial were transcribed from speeches and the latter was also transliterated. In general, we observed that the model's performance has a direct influence on the results of input shortening criteria.

Finally, we observed that for GDI, the mean length of correctly predicted shortened input is 4.9 compared to the 9.5 full length average. In ILI, it is 9.2 compared to 18.5 full length average. In ADI-Vardial, the mean length is 9.02 (compared to 43.02 for full inputs) and 3.58 (compared to 19.89 in full input) in ADI-AOC.

Condition	Description	M	Accuracy
p1	prob(current)>prob(previous)	4454	51.50% (2449)
p2	prob(current)<prob(previous)	4130	47.49% (2257)
p3	prob(current)<prob(next)	4048	44.90% (2134)
p4	prob(current)>prob(next)	4605	55.20% (2624)
l1	predicted label(current) equals predicted label(previous)	4549	52.50% (2496)
l2	predicted label(current) equals predicted label(next)	4628	51.40% (2445)
p1 and l1		4143	50.35% (2393)
p2 and l2		3475	43.37% (2061)
p4 and l1		4354	53.57% (2546)
p1 and p4		4351	53.45% (2540)
p2 and p4		3024	37.00% (1762)

Table 3: Input Shortening Results on GDI with Best Model (Bert-base-cased). M= number of fragments that satisfy the criterion.

Condition	Description	M	Accuracy
p1	prob(current)>prob(previous)	8096	71.24% (6435)
p2	prob(current)<prob(previous)	7842	67.17% (6067)
p3	prob(current)<prob(next)	7799	66.17% (5975)
p4	prob(current)>prob(next)	8285	73.7% (6658)
l1	predicted label(current) equals predicted label(previous)	7975	71.8% (6485)
l2	predicted label(current) equals predicted label(next)	7964	72.44% (6543)
p1 and l1		7975	71.8% (6485)
p2 and l2		7240	66.44% (6001)
p4 and l1		8250	74.1% (6694)
p1 and p4		7946	67.3% (6076)
p2 and p4		7964	72.4% (6543)

Table 4: Input Shortening Results on ILI with Best Model (IndicTransformers). M= number of fragments that satisfy the criterion.

Condition	Description	M	Accuracy
p1	prob(current)>prob(previous)	1266	31.43% (469)
p2	prob(current)<prob(previous)	1261	31.09% (464)
p3	prob(current)<prob(next)	1252	29.8% (445)
p4	prob(current)>prob(next)	1279	32.17% (480)
l1	predicted label(current) equals predicted label(previous)	1288	31.56% (471)
l2	predicted label(current) equals predicted label(next)	1265	32.37% (483)
p1 and l1		1267	32.17% (480)
p2 and l2		1247	31.76% (474)
p4 and l1		1276	33.04% (493)
p1 and p4		1254	31.97% (471)
p2 and p4		1255	30.49% (455)

Table 5: Input Shortening Results on ADI-Vardial with Best Model (Bert-base-cased). M= number of fragments that satisfy the criterion.

Condition	Description	M	Accuracy
p1	prob(current)>prob(previous)	10048	7191(66.51%)
p2	prob(current)<prob(previous)	9486	6577 (60.83%)
p3	prob(current)<prob(next)	9899	6868 (63.52%)
p4	prob(current)>prob(next)	9927	7185 (66.45%)
l1	predicted label(current) equals predicted label(previous)	10226	7167 (66.28%)
l2	predicted label(current) equals predicted label(next)	10125	7198 (66.57%)
p1 and l1		9968	7201 (66.6%)
p2 and l2		9230	6789 (62.79%)
p4 and l1		9874	7172 (66.33%)
p1 and p4		9775	7209 (66.67%)
p2 and p4		9230	5566 (51.47%)

Table 6: Input Shortening Results on ADI-AOC with Best Model (AraBERT). M= number of fragments that satisfy the criterion.