Explainability of machine learning approaches in forensic linguistics: a case study in geolinguistic authorship profiling

Dana Roemling
University of Birmingham
University of Helsinki
d.roemling@bham.ac.uk

Yves Scherrer
University of Oslo
University of Helsinki
yves.scherrer@ifi.uio.no

Aleksandra Miletić
University of Helsinki
University Sorbonne Nouvelle
aleksandra.miletic@helsinki.fi

Abstract

Forensic authorship profiling uses linguistic markers to infer characteristics about an author of a text. This task is paralleled in dialect classification, where a prediction is made about the linguistic variety of a text based on the text itself. While there have been significant advances in recent years in variety classification, forensic linguistics rarely relies on these approaches due to their lack of transparency, among other reasons. In this paper we therefore explore the explainability of machine learning approaches considering the forensic context. We focus on variety classification as a means of geolinguistic profiling of unknown texts based on social media data from the German-speaking area. For this, we identify the lexical items that are the most impactful for the variety classification. We find that the extracted lexical features are indeed representative of their respective varieties and note that the trained models also rely on place names for classifications.

1 Introduction

Forensic authorship analysis is a key area of research within forensic linguistics that seeks to gain information about the author(s) of a text. Generally, there are two central domains of research within authorship analysis: *comparative authorship analysis* uses various methodologies to compare questioned and known documents to attribute authorship, while *sociolinguistic* or *authorship profiling* relies on the analysis of questioned texts alone to infer characteristics of an author (Grant, 2022; Roemling and Grieve, 2024). Both areas of authorship analysis can be approached qualitatively and/or quantitatively if the amount of available data permits it. For example, quantitative work in authorship profiling has researched inferring age or gender (Nini, 2018) or native language influence (Kredens et al., 2019) from questioned documents.

Nevertheless, forensic authorship profiling is often carried out in a manual or qualitative way, relying on the expertise of the forensic linguist. This is evident in famous examples like the *Unabomber* analysis (Leonard et al., 2017) or the *devil strip* ransom note (Shuy, 2001). In both cases, law enforcement was interested in the regional origin of the authors. This background can be inferred through analyzing the regional linguistic variation, i.e., the use of regional dialect, in the questioned documents. This is referred to as *regional* or *geolinguistic profiling* (Roemling and Grieve, 2024) and is a task parallel to inferring the regional variety of a document as is done in language identification (Jauhiainen et al., 2019).

Even though research in forensic linguistics works more and more with statistical and computational approaches (e.g., Bevendorff et al., 2023; Ishihara et al., 2024; Nini et al., 2024), authorship profiling often remains a manual task. This is at times credited to the black-box approaches in current NLP research, meaning that the lack of explainability precludes these approaches from being used in legal settings (see Nini, 2023).

2 Related work

The interest in explainability of machine learning (and in particular, neural) approaches is not only a relevant research area for forensic linguistics. Previous work, including on language identification, has focused on understanding how classifiers come to their predictions. Research started by creating an interpretable model around the actual classification approach to explain the predictions of the classifier (Ribeiro et al., 2016). Li et al. (2017) employed representation erasure to propose a methodology for interpretability research while Jacovi et al. (2018) explored how filters in CNNs can be understood in text classification research, finding that filters do in fact learn different classes of ngrams. Furthermore, Ehsan et al. (2019) showed that by training on hu-

10

man explanation data, models can learn to translate their inner states into understandable explanations. Belinkov et al. (2020) thus summarised the main sub-fields of interest in interpretability research as focusing on "probing classifiers, behavioral studies and test suites, and interactive visualizations".

Xie et al. (2024) were the first to research explainability in a dialect classification context. In order to analyze the classifier, they extracted lexical features that were highly relevant to the classification, aiming to use this knowledge for dialect research and not only as a means to explore machine learning approaches. They relied on lexical items, owing to the complex nature of handling preprocessing like tokenisation or POS-tagging in low(er)-resourced varieties. They indicated that refined approaches would be beneficial. However, previous research on regional variation using social media data has shown that approaches using lexical features provided excellent results (e.g., Doyle, 2014; Huang et al., 2016; Eisenstein, 2017; Grieve et al., 2018, 2019).

Xie et al. (2024) proposed two different approaches, one intrinsic and one post-hoc, to extract features relevant to the dialect classification. In the intrinsic approach, the authors added a local interpretability layer to the dialect classifier which was trained together with the model and output the relevance of a feature for the classification. For the post-hoc approach, Xie et al. used a leave-one-out (LOO) method, where the change in prediction probability if a feature was left out was interpreted as a "relevance score" of that particular feature.

In a forensic authorship profiling setting, an approach like this appears beneficial as it reaches high accuracies in language identification, while similarly providing explanations by extracting the features that influenced the classification. While the original study focused on improving research methods in dialectology, we apply the method to evaluate its usefulness in a forensic context. Additionally, the approach has the advantage of eliminating or at least minimising researcher bias as the models learn the relevant features themselves, whereas it is the forensic linguist's expertise that culls the features in a qualitative analysis (see Grant and Grieve, 2022)[1]. While the approach does not fully explain the inner workings of the model, experts can use the extracted features to a) verify that

the model indeed reached a sound decision, for example by evaluating the features against previous dialectological findings, and b) use the explanations to introduce the method to law enforcement or jurisprudence. Even if the classifiers themselves do not meet court admissibility standards (Coulthard, 2013; Hammel, 2022), extracted features can be used for authorship work to contribute to a more objective analysis.

3 Data

We work with a corpus of German social media data from the platform Jodel. The corpus was collected by Hovy and Purschke (2018); Purschke and Hovy (2019). It has also been used in VarDial classification tasks (Gaman et al., 2020; Chakravarthi et al., 2021). Jodel is structurally similar to Twitter/X, however it only allows anonymous posts. Users of Jodel can interact with other users in a 10-15 km radius around their own location, so all posts are geolocated. The corpus contains posts from Austria, Germany and Switzerland. While most of the data is written in standard German, it shows clear regional patterns. Especially in Switzerland, Austria and, in parts, Bavaria writing is considerably further from standard German (Purschke and Hovy, 2019). Posts from Romandy contain substantial amounts of French. This data differs from the corpus used in the original study (see Xie et al., 2024) in terms of register and genre.

The corpus consists of approximately 240 million tokens from about 8500 locations, however only 388 locations have a token count of over 10k. For our classification experiments, we mapped these locations onto wider dialect regions following three settings with 3, 4 and 5 classes respectively. The 3-class distinction is based on national borders, so the classes reflect Austria, Germany and Switzerland. In the 4-class setting Germany is additionally divided into two parts, north and south (at latitude 50.33° N) and for the 5-class setting the southern region of Germany is further split into east and west (at longitude 9.97° E) (see Figure 1). These divisions were operationalized based on knowledge from traditional dialectology (Wiesinger, 1983; König, 2004).

We randomly sampled 200k posts per class for training, and 20k posts per class for development and testing, respectively. On average, a post contains 11.5 tokens. Besides some simple whitespace normalization (i.e., removing line breaks and tabs
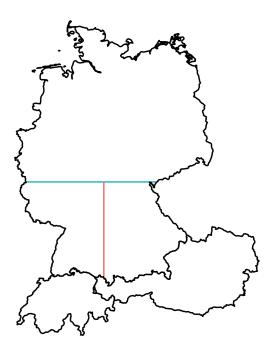
Figure 1: Operationalization of dialect regions

Classes	Random	XLM-RoBERTa	German BERT
3	33%	75.31%	74.91%
4	25%	58.57%	58.44%
5	20%	47.74%	47.64%

Table 1: Classification accuracies on the development sets.

Considering two base models and three settings (3, 4 and 5 classes), our experiments yield six dialect classifiers. The training is done with the `simpletransformers` library[4]. Each model is trained for 10 epochs with a maximum sequence length of 256 (subword tokens) and a batch size of 64 samples. We use default values for all other parameters.

Table 1 shows the classification accuracies of these models on the development sets. It can be seen that all models outperform the random baseline by a large margin. The difference between the two base models is marginal and we thus focus on the German base model.

4.2 Leave-one-word-out classification

The LOO method used by Xie et al. (2024) processes each sentence of the test set independently to detect the words that contribute most to the classification. It consists of the following steps:

1. Select an instance x of the test set, run it through the dialect classifier, and record the predicted class $\hat{y}$ as well as the prediction score ℓ. If the prediction is incorrect ($\hat{y} \neq y$), skip this instance and move to the next one.[5]

2. Select one word of the instance and remove it from the instance (let x_i denote the instance x from which word i is removed), run it through the dialect classifier, and record the prediction score ℓ_i.[6]

3. Measure the impact of the removed word on the classification performance (Δ_i) by subtracting the score of the incomplete instance from the score of the complete instance: $\Delta_i = \ell - \ell_i$. We call Δ_i the *impact score* of word i.

inside a post), we did not perform any preprocessing.

4 Methods and experiments

The basis for our analysis is the post-hoc LOO approach proposed by Xie et al. (2024). While we replicate most of the methodology, we mark any changes from the original in our explanations below. Crucially, and in contrast to the original dialectal research interest, we apply the approach with the goal of evaluating its usefulness in a forensic setting. Additionally, we work with data that is different in terms of register and genre and thus adds to our overall understanding of the approach.

4.1 Dialect classifiers

The approach described by Xie et al. (2024) starts by fine-tuning a BERT-based language model on the dialect classification task. Following Xie et al. (2024), we rely on the multilingual model `xlm-roberta-base`[2] (Conneau et al., 2020) but we also experiment with a base model specifically trained on German data, `dbmdz/bert-base-german-cased`[3].

---

[2] https://huggingface.co/FacebookAI/xlm-roberta-base

[3] https://huggingface.co/dbmdz/bert-base-german-cased

[4] https://simpletransformers.ai/

[5] This corresponds to the *isCorrect* constraint of Xie et al. (2024).

[6] Xie et al. (2024) select the word to be removed by position, with the consequence that if a word occurs several times in the same sentence, only one of its occurrences will be removed at a time. They then only consider the occurrence that produced the biggest difference. We simplify this part by iterating over the set of unique words and removing all occurrences of the selected word at the same time.

4. Repeat steps 2 and 3 for each word of the sentence.

5. Select the 5 words with the highest impact score.[7]

The steps described above produce *explanations* at instance level, i.e., the most impactful words of each instance. In a forensic case setting with limited data, this could already be leveraged by using the impactful features in a qualitative analysis or simply evaluating and explaining the prediction made by the classifier. Consequently, this can be an interesting result in itself, but for the analysis reported here, we aggregate the explanations across all instances of the test set for evaluation. The resulting list is processed in the following way:

1. Words that were selected as explanations for more than one class are eliminated from further consideration.[8]

2. Words that figure as explanations for only one instance are eliminated from further consideration.

3. For each remaining word, we compute the average impact score on the basis of the individual impact scores.[9]

4. For each class, we select the 100 words with the highest average impact scores for the analysis.

5 Results

Following the method described above, we produce a list of 100 words with the highest impact scores per class in all settings. A manual inspection of the lists quickly shows a prevalence of place names and related items like *Zürich* or *Österreicher* (G 'Austrian')[10]. Therefore, as a first step, we count the amount of these words among the top 100 and find that, on average, 14% of words are local references. In terms of classification, these results are expected (see, e.g., Nasar et al., 2022) and it is apparent how these words are indicative of location given what they denote. Although it does not take a forensic linguist to understand the connection of these items

to location, it is noteworthy that the classification models pick up on these words and that they match the region they are impactful for.

Given that the results are similar for classes based on country borders across the three settings, we focus our analysis on the 5-class setting. As a reminder, in this setting, Austria and Switzerland form individual classes, whereas Germany is split into three regions (see Figure 1). Generally, the results show that a large proportion of extracted words are regionalized and some are prototypical dialect items.

For Switzerland, we find both Swiss German and French lexical items to be the most impactful, such as *Kei* (SWG 'not a/no'), *Isch* (SWG 'is') and *bim* (SWG 'at the'), or *pourquoi* (FR 'why'), *Avec* (FR 'with') and *raison* (FR 'reason'). Whereas for Austria, the data shows that some of the most impactful lexical items are *Oasch* (AG 'ass'), *Gspusis* (AG 'affairs'), and *Matura* (AG 'high-school diploma'). These three items are examples of a textualization of regional pronunciation, a regional item and a standard Austrian variant, respectively. Also, items such as *Jus* (AG 'law (studies)'), which are relevant to the Jodel demographic of mostly students under the age of 27, are extracted. This indicates that an automated classification between the three countries, for instance to clarify jurisdiction, seems reasonable.

For the three German classes we find that a large proportion of top words appear textualized in standard German as opposed to more colloquial spellings including abbreviations and ellipses, which we may expect. Examples of this include *Dankeschön* (G 'thank you') or *Vorname* (G 'first name'). Considering the division within Germany, we find that several forms of the verb *gucken* (G 'to look/watch') are impactful for the northern class, which is a variant we know to be regionalized and appearing in varieties in central and northern Germany (König, 2004, p. 235). For the south-east class the data shows items like *Ritter* (G 'knight'), *#traudel* (female first name) and local beer types. For the south-west we find that items identified as relevant by Purschke and Hovy (2019), are also extracted by the LOO model, like *Möppes* (G 'breasts' or 'female user') and *Lörres* (G 'penis' or 'male user'), although the authors argue that these forms are more Jodel- than region-specific. These items are also impactful for the Germany-class in the 3-class setting.

---

[7]Xie et al. (2024) omit this step in the description of their work, but it is present in their code.

[8]This corresponds to the *isUnique* constraint of Xie et al. (2024).

[9]Xie et al. (2024) use TF-IDF to rank the words. We find that the simpler approach of averaging the scores is sufficient for our purposes.

[10]Throughout the remainder of the paper, the examples from Germany are marked with *G*, those from Austria with *AG*, and those from Switzerland with *SWG*. Examples from Romandy are marked *FR* for French.

6 Conclusion

In this paper, we employed Xie et al. (2024)'s approach as a means of geolinguistic profiling to understand how a method like this could be applied in a forensic context given its explainability. We have found that the dialect classifiers outperform the random baseline by a large margin in all settings, but that accuracy decreases for settings with more closely-related classes. While we recognize that in forensic contexts the focus needs to be on false predictions and hard-to-classify cases, this paper considers the explainability of the approach. To this end we have found that the LOO model extracts meaningful regional features reflecting the variety used in the corresponding area. On average, 14% of extracted features are place names or similar items. While an analysis for a place name like *Wien* does not need a forensic linguist, extracting features that are not based on a linguist's expertise is a valuable contribution of this approach even if it is not directly used for automated classification.

Limitations

For this paper we have worked with the geolocation of the Jodel posts as the gold label for the dialect regions used as classes in the classification task. However, there is noise in this data as people move and use varieties from different regions in the same place. Further analysis of the incorrect classifications may allow us to identify these instances.

For further work it may be beneficial to remove non-German varieties before training. Additionally, given the high percentage of place names and related lexical items, preprocessing to remove named entities (see, e.g., Darji et al., 2023) may help focus the extraction on dialectal lexical items.

Acknowledgements

References

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, page 1–5, Online. Association for Computational Linguistics.

Janek Bevendorff, Mara Chinea-Ríos, Marc Franco-Salvador, Annina Heini, Erik Körner, Krzysztof Kredens, Maximilian Mayerl, Piotr Pęzik, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, and Eva Zangerle. 2023. *Overview of PAN 2023: Authorship Verification, Multi-author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection: Extended Abstract*, volume 13982 of *Lecture Notes in Computer Science*, page 518–526. Springer Nature Switzerland, Cham.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 5454–5476.

Bharathi Raja Chakravarthi, Mihaela Gaman, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nicola Ljubešić, Niko Partanen, Ruba Priyadharshini, and Christoph Purschke. 2021. Findings of the vardial evaluation campaign 2021. The Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Malcolm Coulthard. 2013. On admissible linguistic evidence. *Journal of Law and Policy*, 21(2):441–466.

Harshil Darji, Jelena Mitrović, and Michael Granitzer. 2023. German bert model for legal named entity recognition. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, page 723–728. ArXiv:2303.05388 [cs].

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, page 98–106, Gothenburg, Sweden. Association for Computational Linguistics.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User*

Interfaces, page 263–274, Marina del Ray California. ACM.

Jacob Eisenstein. 2017. *Identifying Regional Dialects in On-Line Social Media*, 1 edition, page 368–383. Wiley.

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. *Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects*, page 1–14.

Tim Grant. 2022. *The Idea of Progress in Forensic Authorship Analysis*, 1 edition. Cambridge University Press.

Tim Grant and Jack Grieve. 2022. *The Starbuck case: Methods for addressing confirmation bias in forensic authorship analysis*, first edition edition, page 13–28. Wiley Blackwell, NJ & West Sussex.

Jack Grieve, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019. Mapping lexical dialect variation in british english using twitter. *Frontiers in Artificial Intelligence*, 2:11.

Jack Grieve, Andrea Nini, and Diansheng Guo. 2018. Mapping lexical innovation on american social media. *Journal of English Linguistics*, 46(44):293–319.

Andrew Hammel. 2022. *Linguistic Expert Evidence in the Common Law*, page 55–84. Springer International Publishing, Cham.

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*, page 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.

Shunichi Ishihara, Sonia Kulkarni, Michael Carne, Sabine Ehrhardt, and Andrea Nini. 2024. Validation in forensic text comparison: Issues and opportunities. *Languages*, 9(2):47.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, page 56–65, Brussels, Belgium. Association for Computational Linguistics.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Krzysztof Kredens, Ria Perkins, and Tim Grant. 2019. Developing a framework for the explanation of interlingual features for native and other language influence detection. *Language and Law=Linguagem e Direito*, 6(2):10–23.

Werner König. 2004. *Dtv-Atlas deutsche Sprache*, 14., durchgesehene und aktualisierte edition. Deutscher Taschenbuch Verlag, München.

Robert Leonard, Juliane E. R. Ford, and Tanya K. Christensen. 2017. Forensic linguistics: Applying the science of linguistics to issues of the law. *Hofstra Law Review*, 45(3):881–898.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Understanding neural networks through representation erasure. (arXiv:1612.08220). ArXiv:1612.08220 [cs].

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2022. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys*, 54(1):1–39.

Andrea Nini. 2018. Developing forensic authorship profiling. *Language and Law / Linguagem e Direito*, 5(2):38–58.

Andrea Nini. 2023. *A Theory of Linguistic Individuality for Authorship Analysis*, 1 edition. Cambridge University Press.

Andrea Nini, Oren Halvani, Lukas Graner, Valerio Gherardi, and Shunichi Ishihara. 2024. Authorship verification based on the likelihood ratio of grammar models. (arXiv:2403.08462). ArXiv:2403.08462 [cs].

Christoph Purschke and Dirk Hovy. 2019. Lörres, möppes, and the swiss. (re)discovering regional patterns in anonymous social media data. *Journal of Linguistic Geography*, 7(22):113–134.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. (arXiv:1602.04938). ArXiv:1602.04938 [cs, stat].

Dana Roemling and Jack Grieve. 2024. Forensic authorship analysis. *CREST Security Review*, 18: Communication.

Roger W. Shuy. 2001. Dare's role in linguistic profiling. *Dictionary of American Regional English Newsletter*, 4(33):1–5.

Peter Wiesinger. 1983. *Die Einteilung der deutschen Dialekte*, page 807–900. de Gruyter, Berlin, New York.

Roy Xie, Orevaoghene Ahia, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Extracting lexical features from dialects via interpretable dialect classifiers. In *Proceedings of the 2024 Conference of the North*