Master's Thesis

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Arts in Computational Linguistics

Explainable Machine Learning in Linguistics and Applied NLP:

Two Case Studies of Norwegian Dialectometry and Sexism Detection in French Tweets

Author:
Verena Blaschke

Supervisors:
Dr. Çağrı Çöltekin
Prof. John Nerbonne

Seminar für Sprachwissenschaft
Eberhard-Karls-Universität Tübingen

26 May 2021

Hiermit versichere ich, dass ich die Arbeit selbständig verfasst, keine anderen als die angegebenen Hilfsmittel und Quellen benutzt, alle wörtlich oder sinngemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet habe und dass die Arbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen ist und dass die Arbeit weder vollständig noch in wesentlichen Teilen bereits veröffentlicht wurde sowie dass das in Dateiform eingereichte Exemplar mit den eingereichten gebundenen Exemplaren übereinstimmt.

I hereby declare that this paper is the result of my own independent scholarly work. I have acknowledged all the other authors' ideas and referenced direct quotations from their work (in the form of books, articles, essays, dissertations, and on the internet). No material other than that listed has been used.

Tübingen, 26 May 2021

Verena Blaschke

Abstract

This thesis presents an exploration of explainable machine learning in the context of a traditional linguistic area (dialect classification) and an applied task (sexism detection). In both tasks, the input features deemed especially relevant for the classification form meaningful groups that fit in with previous research on the topic, although not all such features are easy to understand or provide plausible explanations. In the case of dialect classification, some important features show that the model also learned patterns that are not typically presented by dialectologists. For both case studies, I use LIME (Ribeiro et al., 2016) to rank features by their importance for the classification. For the sexism detection task, I additionally examine attention weights, which produce feature rankings that are in many cases similar to the LIME results but that are over all worse at showcasing tokens that are especially characteristic of sexist tweets.

Contents

List of Figures

List of Tables

List of Abbreviations

Machine learning and NLP

BERT	Bidirectional encoder representations from transformers
BPE	Byte pair encoding
FFNN	Feed-forward neural network
LIME	Local interpretable model-agnostic explanations
LSTM	Long short-term memory
ML	Machine learning
MLP	Multi-layer perceptron
NLP	Natural language processing
RNN	Recurrent neural network
SVM	Support vector machine
TF-IDF	Term frequency–inverse document frequency

Tokenization

`<SOS>`	Start of sequence
`<EOS>`	End of sequence
`<SEP>`	Separator
`-`	Middle of word
`</w>`	End of word

Linguistic glosses

ACC	Accusative
ADJ	Adjective
FEM	Feminine
NOM	Nominative
PCP	Participle
PRET	Preterite
PST	Past
SG	Singular
PL	Plural

Other

V	Vowel

Chapter 1

Introduction

Many frequently used machine learning (ML) models for text classification are black-box models that are often very good at identifying relevant patterns in their input data, allowing them to be highly accurate classifiers. However, using these models, we can easily test how good they are at classifying some given input data, but extracting what information they learned is less trivial.

Exploring which patterns a classifier has learned and how they affect its predictions is inherently interesting, even (or especially) in two very different scenarios. In more traditional linguistic contexts, explainable machine learning allows us to explore which properties of a dataset for a given task black-box machine learning models find significant, and to what extent they overlap (or do not overlap) with what experts in that field have found with the help of traditional, non-ML methods. I examine this with in a case study of Norwegian dialect classification.

In applied cases, exploring on what basis an ML model makes its decisions allows us to find out if the model is learning patterns that are desired and plausible to humans, rather than spurious or even unwanted correlations that might not generalize well to new data or be actively harmful. In my second case study, I explore this in the context of detecting sexist content in French tweets.

This thesis is structured as follows: I first introduce the explainable machine learning techniques that I use (chapter 2). In chapter 3, I describe the dialectometric case study, and in chapter 4, I present the case study involving tweets. I conclude the thesis in chapter 5.

The code for all experiments can be found at `https://github.com/verenablaschke/ma-thesis/releases/tag/ma-thesis`.

Chapter 2

Explainable machine learning

Barredo Arrieta et al. (2020) present an overview of recent research on explainable machine learning. They distinguish between three types of methods for arriving at these explanations: using transparent models in which the feature interactions are relatively easy to understand for a human (such as decision trees or linear regression models), using model-agnostic post-hoc methods (such as creating local explanations), and using post-hoc methods that are specific to the model architecture.

In this thesis, I use two types of explainable machine learning: LIME, which is a model-agnostic post-hoc approach, and attention weights in neural models, which are specific to that architecture. In the first part of this chapter, I introduce LIME, which I use in both case studies (subsection 2.1.1). In the second part (section 2.2), I illustrate how attention layers in neural models work and summarize the discussion on whether they can be used for explaining model decisions. I use attention in the sexism detection case study.

2.1 Local Interpretable Model-agnostic Explanations

2.1.1 Local explanations with LIME

One popular explanation technique is LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016). This technique works on an instance level. Given an input instance and a trained classifier, LIME fits an interpretable model that makes similar predictions on a level local to this specific input.

The following explains how LIME works, based on the description by Ribeiro et al. and their implementation of the algorithm.[1] Where details differ for text classification tasks and other applications, I describe what applies to text classification models. Moreover, where the authors' implementation differs from the description in their article, I describe the implemented version, as this is what my approach is directly based on. The exact version of the code that I use can be found at `https://github.com/verenablaschke/ngram_lime/releases/tag/ma-thesis`. Figure 2.1 illustrates the approach.

Any given input to a model f is transformed into an input representation $x \in \mathbb{R}^d$, e.g. a matrix containing word embeddings. The model can use this input to produce a probability distribution over labels, $f(x)$. An interpretable version of this is a binary vector $x' \in \{0,1\}^{d'}$ that indicates the presence or absence of discrete, human-understandable features on which x is based; for instance, which words of the vocabulary in the training data are present in this sample. The function m denotes converting the explainable feature vector x' into the input representation x for the machine learning model: $m(x') = x$.

To explore the contribution of each of the non-zero features in x', LIME samples instances from this vector's neighbourhood. Randomly changing some of the ones in x' to zeroes produces a perturbed sample $z' \in \{0,1\}^{d'}$, from which the model input $z \in \mathbb{R}^d$ can be inferred. The model's predicted label distribution for z is $f(z)$, and the probability associated with each class $c \in C$ is indicated by $f_c(z)$.

Explanations are derived at a class level, rather than encompassing the entire label distribution at once. An explanation $g_c \in G_c$ is an interpretable model, where G_c is a set of potential sparse linear models such that $g_c(z') = w_{g_c} \cdot z'$. These models are Ridge regression models that attempt to predict $f_c(z)$. An explanation model's complexity $\Omega(g_c)$ is its number of non-zero feature weights. Minimizing the complexity thus favours models that only focus on a small set of features and are therefore easier to understand for humans. Additionally, it is possible to explicitly set the upper bound of $\Omega(g_c)$. The final choice of the explanation is determined by solving the following:

$$\xi(x) = \underset{g_c \in G_c}{\arg\min}\, \mathcal{L}(f, g_c, \pi_{x'}) + \Omega(g_c) \tag{2.1}$$

where $\mathcal{L}$ measures how dissimilar the original model's predictions are from the output of g_c. The difference between the prediction distributions is weighted by the proximity

[1] The code is available at `https://github.com/marcotcr/lime`; last accessed March 10th, 2021 (commit `a2c7a6f`).

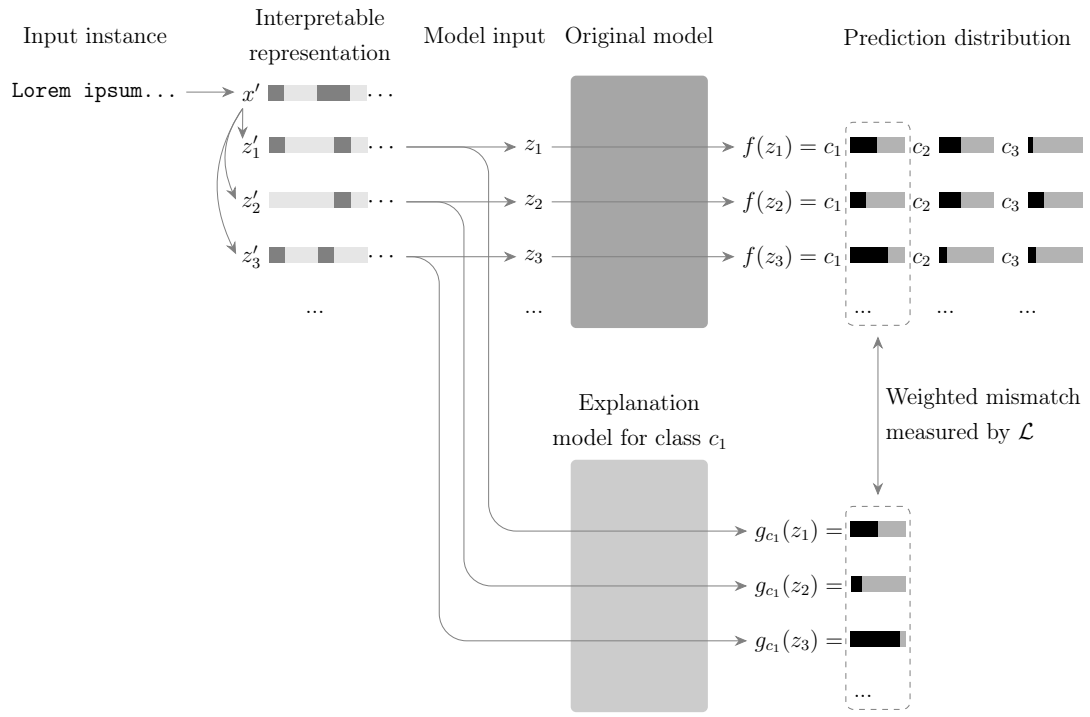


Figure 2.1: LIME generates samples in the neighbourhood of a given instance and compares the predictions of the model to be explained to that of a potential explanatory model.

between x' and z', $\pi_{x'}(z')$, to ensure that g is locally faithful to f:

$$\mathcal{L}(f, g_c, \pi_{x'}) = \sum_{z,z' \in Z} \pi_{x'}(z')(f_c(z) - g_c(z'))^2 \tag{2.2}$$

The proximity between x' and z' is based on the cosine distance between these binary vectors:

$$\pi_{x'}(z') = \exp(\frac{-(1 - \frac{x'z'}{\|x'\|\|z'\|})^2}{\sigma^2}) \tag{2.3}$$

where the kernel width σ is set to 25 in the case of text classification models. Experiments by Garreau and von Luxburg (2020) show that a poor choice of the kernel width σ can lead to important features being ignored by the explanation model, but they are not aware of a good heuristic for picking σ.

Extracting the weight matrix from the model chosen in Equation 2.1 produces importance values for the interpretable features encoded by x'.

Garreau and von Luxburg (2020) provide a theoretical analysis of LIME and find that these importance scores are proportional to the original model's partial derivatives at x. Such local gradients have also been proposed as an approach of explaining machine learning decisions at an instance level, as in work by Baehrens et al. (2010).

2.1.2 Global explanations with LIME

In addition to introducing LIME as a means to generate explanations for individual input instances, Ribeiro et al. (2016) also propose a way of gaining global insights into the feature importance scores. Assuming X is a set of instances that represents the data on which the original model f is to be used, then for each $x \in X$, it is possible to fit an interpretable model according to Equation 2.1 and retrieve the local explanation. The explanations can then be stored in an explanation matrix $\mathcal{W}_{|X|\times d'}$, such that each row corresponds to an input instance and each column contains the weights associated with one of the interpretable feature representations. A global importance score for a feature in column j is then

$$\text{importance}_{\text{Ribeiro}}(j) = \sqrt{\sum_{i=1}^{|X|} |\mathcal{W}_{ij}|} \tag{2.4}$$

Note that this is based on the absolute values of the individual importance scores. Accordingly, a high global importance score can mean that the given feature is a strong positive predictor for a given class, or that it is a strong negative predictor.

Since I am specifically interested in the predictors per label, I instead define the global importance score of a feature as the mean of all of its local importance scores:

$$\text{importance}_{\text{mean}}(j) = \frac{\sum_{i=1}^{|X|} \mathcal{W}_{ij}}{|X|} \tag{2.5}$$

This latter method is also used by Garreau and von Luxburg (2020) in their analyses of LIME.

Garreau and von Luxburg point out that the error of the local explanation can be viewed as an indicator of the explanation's quality. In an additional experiment, I scale each individual local importance score in $\mathcal{W}$ by how close the local model's predictions are to the original model's predictions, before plugging it into Equation 2.4 and Equation 2.5. This gives more weight to coefficients from reliable local models and less weight to scores from models that do not fit the data well. I use the coefficient of determination of the prediction (R^2) as the weight for scaling if it is positive, and a weight of 0 otherwise. This gives an upper bound of 1.0 to the weight (if the interpretable model perfectly emulates the original model's predictions locally) and a lower bound of 0.0 (if the local model's predictions are not better than always guessing the expected value of $f(z)$). However, in preliminary experiments for both the dialect labelling task and the tweet classification task, scaling the scores in this way does not affect the outcome much (neither the ranking of the features with the highest global importance scores, nor the relationship between importance scores and representativeness and distinctiveness, which are introduced in subsection 2.1.4). In the case of the dialect data, the average R^2 score of the local models is 0.87, with a standard deviation of 0.09. The R^2 score for the tweet classification is 0.89, with a standard deviation of 0.07.

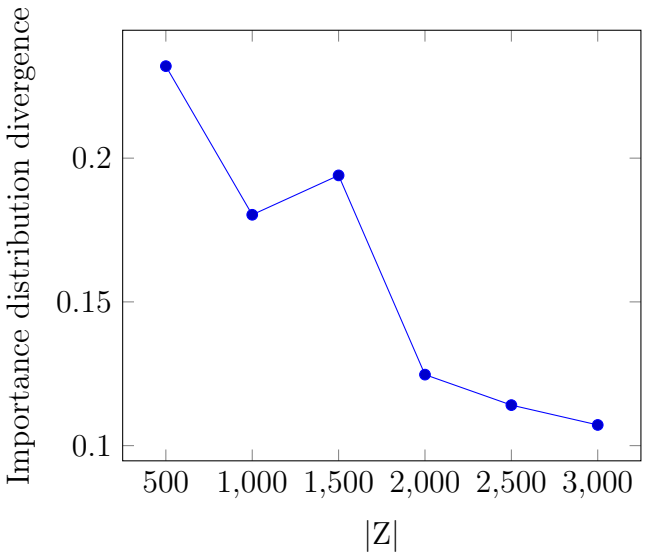


Figure 2.2: The average divergence of importance coefficient distributions across ten runs of LIME for different numbers of input feature permutations Z of the Twitter data.

2.1.3 LIME settings

Garreau and von Luxburg find that, if the number of neighbourhood samples is large enough, the explanation model's feature coefficients tend to be stable across different runs of LIME. To generate explanations that are consistent across different runs of LIME, I experiment with varying the number of input permutations per LIME sample instance, $|Z|$.

To measure the stability, I train one SVM as classification model on 90 % of the available data and reserve the remaining instances as test data. For each value of $|Z|$ that I test, I use the trained SVM and ten initializations of LIME to calculate global feature importance scores. To quantify the distance between the ten resulting importance score distributions, I calculate the mean of the Jensen-Shannon distance (Lin, 1991) between each (non-identical) pair of initializations.

As shown in Figure 2.2, the divergence between importance score distributions noticeably drops when $|Z| \geq 2000$ for the Twitter data. This also happens when $|Z| \geq 1000$ for the dialect classification task (albeit less clearly; see Figure 2.3)

I set the maximum number of features with non-zero coefficients per utterance explanation, $\Omega(g_c)$, to 100.

2.1.4 Representativeness and distinctiveness

In order to further inspect how informative a feature is, I use the measures of *representativeness* and *distinctiveness*, which are based on the metrics of the same name used by Wieling and Nerbonne (2011) in the context of investigating features

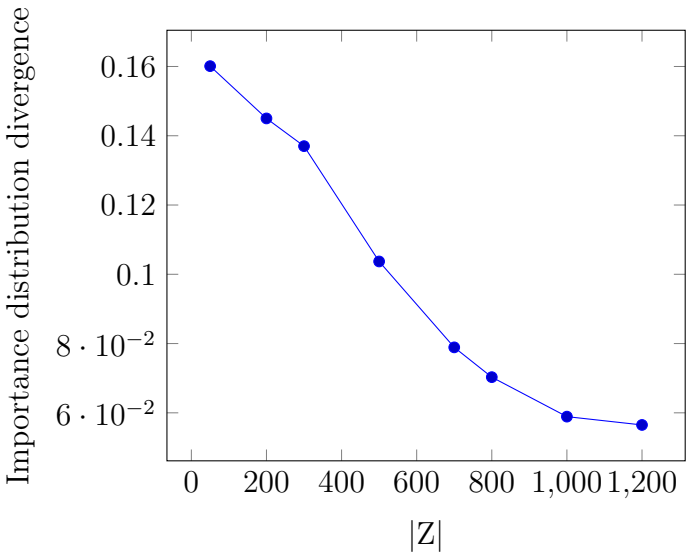


Figure 2.3: The average divergence of importance coefficient distributions across ten runs of LIME for different numbers of input feature permutations Z of the dialect data.

in dialectometry tasks. *Distinctiveness* is also similar to the *polarized weirdness index* that Poletto et al. (2020) use for analyzing hate speech corpora.

In the following, I define X as the set of (test) input data instances, $label(x)$ as the function returning the gold standard label for a given instance of the data, and $features(x)$ as the function returning the set of explainable features contained in x (which is encoded by LIME's x'). *Representativeness* measures the proportion of instances containing a given feature f within the set of utterances that have a given gold standard label l:

$$\text{representativeness}(f, l) = \frac{|\{x \in X \mid f \in \text{features}(x), \text{label(x)} = l\}|}{|\{x \in X \mid \text{label}(x) = l\}|} \tag{2.6}$$

Complementarily, *distinctiveness* measures the proportion of instances containing a given gold standard label l within the set of utterances that have a given feature f, normalized by the relative size of the label class l:

$$\text{relative-size}(l) = \frac{|\{x \in X \ \text{label}(x) = l\}|}{|X|} \tag{2.7}$$

$$\text{relative-occurrence}(f, l) = \frac{|\{x \in X \mid f \in \text{features}(x), \text{label}(x) = l\}|}{|\{x \in X \mid f \in \text{features}(x)\}|} \tag{2.8}$$

$$\text{distinctiveness}(f, l) = \frac{\text{relative-occurrence}(f, l) - \text{relative-size}(l)}{1 - \text{relative-size}(l)} \tag{2.9}$$

The highest possible distinctiveness score is 1, in which case the given feature only occurs in samples with the given label. A distinctiveness score of 0 indicates that

the feature and label co-occur as often as would be expected if they were randomly, independently distributed, i.e. they are independent of one another. Distinctiveness has no fixed (label-independent) lower bound, but negative scores indicate that a feature and a label tend to specifically *not* co-occur.

2.2 Attention

2.2.1 Attention layers in neural networks

I use a neural network with an attention layer as an additional classifier for the tweet classification task. The architecture is based on the ones by Yang et al. (2016) and Sun and Lu (2020). Similar architectures have been used for offensive language detection by Chakrabarty et al. (2019) and Risch et al. (2020), and by Jain and Wallace (2019) in their discussion of attention and explanation. The main difference between my neural model architecture and the others mentioned here, is that I use a feed-forward neural network (FFNN) rather than a recurrent one (RNN). This decision is motivated by the discussion in subsection 2.2.3.

Each model input instance is represented as a sequence of T tokens in an embedding matrix $z \in \mathbb{R}^{T \times e}$. The encoder (in this case a feed-forward neural network) uses this embedding matrix to produce an encoded representation $h \in \mathbb{R}^{T \times m}$.

This representation can be compared to a *context vector* $v \in \mathbb{R}^m$ via a similarity function ϕ to get a distribution of attention weights $\alpha \in \mathbb{R}^T$, one per input token representation:[2]

$$\alpha = softmax(\phi(h, v)) \tag{2.10}$$

These attention weights are what I analyze in section 4.4.

The context vector is randomly initialized and not connected to the encoder or decoder output. However, it plays a similar role to the *query* in sequence-to-sequence models with attention or in self-attention[3] layers, where that is a representation of the previous timestep by the decoder or encoder, respectively (Bahdanau et al., 2015; Vaswani et al., 2017).[4]

There are several different similarity functions that are widely used for attention architectures. I use the similarity function for *scaled*[5] *dot-product attention*, as introduced by Vaswani et al. (2017):

$$\phi(h, v) = \frac{hv}{\sqrt{m}} \tag{2.11}$$

[2]Yang et al. (2016) introduce an intermediary layer between the encoder and the attention: $u = tanh(Wh + b)$, and compare u (rather than h) to v in Equation 2.10. Sun and Lu (2020) and Jain and Wallace (2019) omit this additional step and proceed as outlined above.

[3]Chakrabarty et al. (2019) found that attention based on context vectors generally yields better results for offensive language detection tasks than self-attention.

[4]Bahdanau et al. (2015) base the query for the first timestep of a sequence on part of the hidden encoder representation for that same timestep.

[5]Normalizing the matrix product based on the hidden layer size before applying the softmax function yields less extreme softmax values. The closer the output of the softmax function is to 0 or 1, the smaller the gradients get, making it harder to efficiently update the model weights during training.

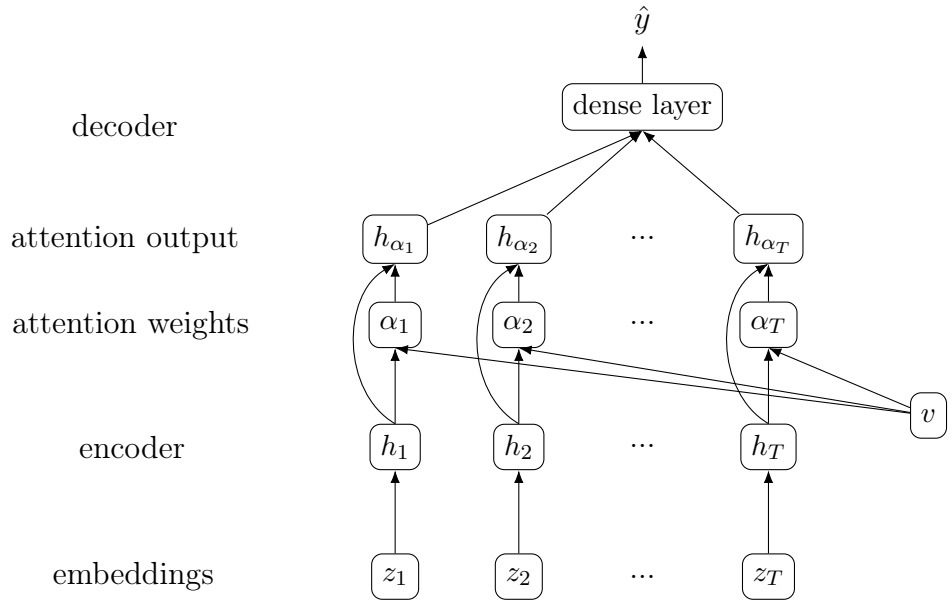


Figure 2.4: The neural classification model encodes token embeddings z with a neural network. Multiplying the resulting matrix h with the context vector v creates token-wise attention weights α, based on which the weighted token representations h_α are generated.

The attention weights are used to produce a representation of the input sequence wherein the individual token representations are weighted by the attention distribution:

$$h_\alpha = \sum_{t=1}^{T} \alpha_t \cdot h_t \tag{2.12}$$

This attention output h_α is then passed to a final network layer (decoder) that generates the predicted label distribution.

2.2.2 Attention weight entropy

Where LIME's loss function limits the number of features per utterance that receive non-zero importance scores, there is no such restriction when the attention weights are calculated. I calculate the *entropy* of each utterance's attention weight vector to determine how informative this attention distribution is:

$$\mathrm{H}(\alpha) = -\sum_{t=1}^{T} \alpha_t \log(\alpha_t) \tag{2.13}$$

In the most uninformative case, where adding the attention layer does not have an impact on the decoder input, each entry within α is $\frac{1}{T}$. This gives the upper bound

for the entropy: $-\log(T)$.

If an utterance contains only one relevant token (per the attention distribution), the attention weight vector is one-hot encoded and the corresponding entropy is $\log(1) = 0$, which is the lower bound for the potential entropy scores.

2.2.3 Attention and explanation

In the past few years, there has been much discussion around whether attention weights are suitable for explaining model predictions. The initial rationale for using attention as a proxy of explanation is easy to see: after all, the attention layer produces a representation of the model input that is weighted in such a way that some input tokens may have a larger influence on the output prediction than others. Whether or under what conditions this can be used for explaining model decisions has been the topic of much discussion. In this section, I summarize the major arguments against and for interpreting attention as explanation as well as common caveats.

Jain and Wallace (2019) examine the merit of analyzing attention weights by carrying out a series of experiments with a neural model containing a bidirectional LSTM (bi-LSTM) followed by an attention layer for different sequence classification tasks.[6] They argue that attention weights fail to be useful as explanations in two ways: they do not consistently correlate with other measures of feature importance and it can be possible to change the learned attention weights with only minor impact on the model predictions.

Correlation with other measures

Jain and Wallace (2019) reason that attention weights should be correlated with other measures of feature importance, such as gradient-based measures or leave-one-out scores. They find that the correlation between gradient-based measures or leave-one-out importance measures and attention weights in bi-LSTMs depends on the choice of dataset and is only in some cases statistically significant. By contrast, the correlation between leave-one-out scores and gradients in the bi-LSTM model is stronger than between attention weights and either of the two other importance measures. However, when training a model whose encoder is a feed-forward neural network instead of a bi-LSTM (Figure 2.5b), the authors find that there is a strong correlation between attention weights and gradients.

Comparison to gradient-based importance rankings

Serrano and Smith (2019) also use gradient-based importance measures as a point of comparison. They base their experiments on different neural models that contain a recurrent or convolutional layer and whose last layer before the decoder is an

[6] The authors also try out different similarity functions for calculating the attention weights (Equation 2.11 in subsection 2.2.1), but find that the choice of similarity function does not make any significant difference.

attention layer. The authors compare importance rankings produced (1) randomly, (2) by attention weight magnitude, (3) based on the decision function's gradient with respect to the attention weight, and (4) by multiplying the gradient with the attention weight. They then remove one input token after the other, in order of descending importance, until the predicted class for the instance changes. The authors find that while removing inputs on the basis of the attention-based ranking produces label changes quicker than when removing inputs randomly, both gradient-based rankings require removing fewer inputs for a label change than the attention-based ranking, indicating that attention alone is not sufficient for uncovering minimal sets of inputs that are the most relevant for the final label prediction.

Serrano and Smith repeat this with models that use feed-forward layers instead of recurrent or convolutional layers. They find that, independently of the ranking approach used, it is sufficient to remove much smaller sets of inputs in order to change the model's predicted class for a given instance.[7] The more context is shared between input representations before the attention layer, the less clearly do the rows in h represent the input tokens with the same indices.

Modifying attention weights on a per-instance basis

Jain and Wallace (2019) argue that changing a model's attention weights for a given input[8] should lead to predicting a different label distribution. They explore this in two experiments:

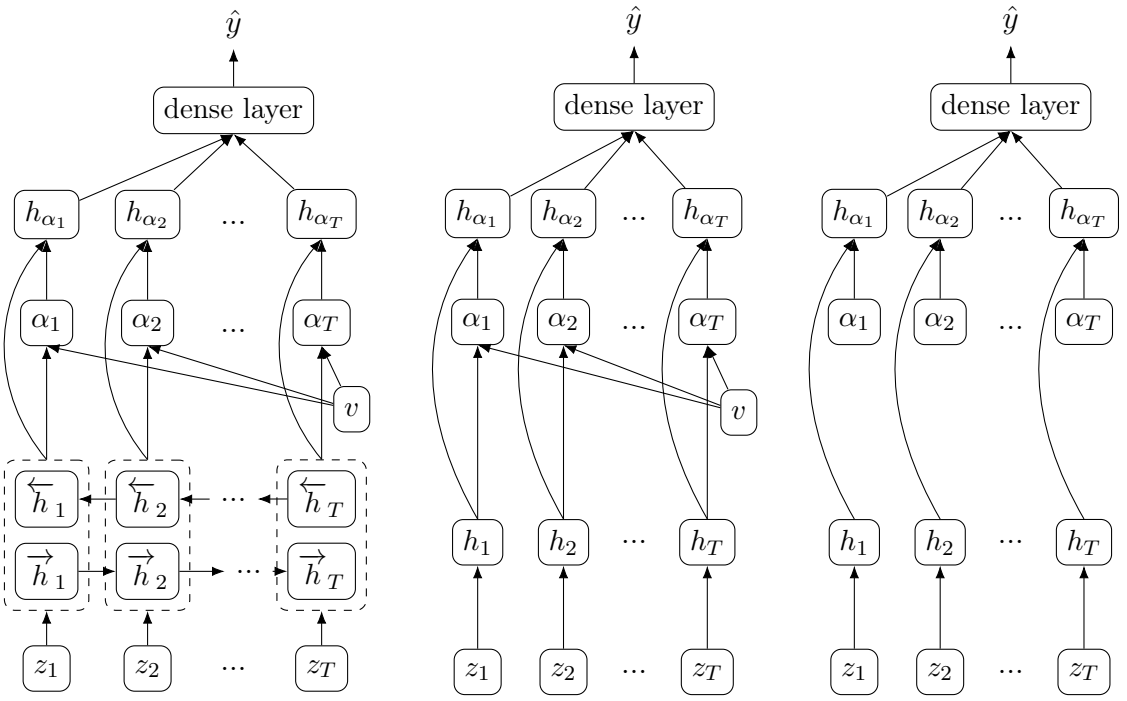
In the first, they calculate the attention weights for a given input as usual, but then randomly permute the entries in the attention weight vector before calculating the attention-weighted decoder input (Equation 2.12 in subsection 2.2.1). The authors find that, while the results also change somewhat from dataset to dataset, there are many cases where permuting an attention weight vector with low entropy (i.e. a distribution that implies that only few input tokens are relevant) does not result in a markedly different prediction.

In the second experiment, Jain and Wallace modify the attention weights such that their distribution is as different as possible from the original attention weight distribution while still yielding very similar label predictions. They observe that in many cases, it is indeed possible to find such *adversarial attention weights* that imply different importance assignments to the input tokens.

Jain and Wallace also find that in some tasks, whether the prediction changes or not depends on the (originally) predicted label: shuffling the attention weights matters very little (or finding adversarial attention weights is not possible) when the originally predicted label belongs to one class, but permuting the weights leads to very different

[7] In general, using a feed-forward layer yields smaller minimal input sets than using a convolutional layer, which in turn results in smaller minimal input sets than when using a recurrent layer.

[8] Wallace (2019) reasons that focusing on instance-level attention weights rather than the model-wide context vector from which the attention weights are calculated makes sense as attention is often used when seeking to explain the prediction for individual instances. However, this is a moot point in the context of this thesis, as I consider aggregated attention weights.



(a) Bidirectional LSTM with learned attention.

(b) FFNN with learned attention.

(c) FFNN with imposed attention.

Figure 2.5: Three model architectures used in experiments by Jain and Wallace (2019) (subfigures 2.5a and 2.5b) and Wiegreffe and Pinter (2019) (all subfigures). The set-up in 2.5b is identical to Figure 2.4. The architecture in 2.5c differs from the first two in that its attention weights are frozen and not trained with the model.

predictions (or adversarial attention exists) when the unmodified model predicts a different class.

Serrano and Smith (2019) also carry out an experiment on instance-level attention weights, although they stress that they focus on "the importance of intermediate quantities, which may themselves already have changed uninterpretably from the model's inputs" after having already been modified by other model layers. In their experiment, they investigate how the predictions change when one of the weights within the attention vector is set to zero and the remaining weights are re-normalized (such that they also sum up to 1). For a given instance, they (separately) remove the highest of the attention weights and a randomly chosen weight in this way and compare how the output distribution over labels changes. The authors find that the larger the difference in the attention weight magnitude between the two removed attention weights is, the more the output distributions tend to diverge, showing that in cases where the attention weight vector is nearly one-hot encoded (i.e. only one entry in h receives nearly all of the attention), removing the input associated with the highest attention weight has a clear impact on the prediction.

Modifying attention weights on a per-model basis

Wiegreffe and Pinter (2019) criticize that experiments involving manipulations of the attention weights treat the attention weights as independent of the model (when the context vector on which the original attention weights are based is a product of the model training as a whole) and that the adversarial weights were created on a per-instance rather than per-model basis.

In a series of experiments, Wiegreffe and Pinter compare the performance of different model architectures with attention layers:

1. *Trained multi-layer perceptron (MLP)*: A feed-forward neural network with attention weights that are learned during training, similar to the architecture described in subsection 2.2.1 but with an FFNN instead of the bi-RNN (Figure 2.5b).
2. *Uniform*: This architecture is similar to the trained MLP, but instead of training attention weights, they are fixed to a uniform distribution ($\alpha = (\frac{1}{T}, \frac{1}{T}, ..., \frac{1}{T})$) (Figure 2.5c).
3. *Base LSTM*: This model also uses frozen weights (Figure 2.5c), but rather than being uniform, they are extracted for each input instance from a bi-LSTM with attention (Figure 2.5a) trained on the same data.
4. *Adversary*: This architecture resembles the trained MLP, but it is trained to make the same predictions as a different attention-based model while creating a context vector that produces maximally different attention weights for each instance.

For all of the tested datasets, the trained MLP outperforms the uniform model, and the model with the pretrained bi-LSTM's attention weights outperforms the trained MLP. For all datasets save one, the adversary model performs much worse than the trained MLP. Wiegreffe and Pinter conclude that "adversarial distributions, even those obtained consistently for a dataset, deprive the underlying model from some form of understanding it gained over the data, one that it was able to leverage by tuning the attention mechanism towards preferring 'useful' tokens." However, while there is a trade-off between the similarity of the predictions and the difference of the attention weights, it is less strong than it would have to be for the original attention to clearly not be manipulable.

Chapter 3

Case study: Norwegian dialect disambiguation

The Norwegian language presents an interesting case for dialectologists in that there does not exist a standard version of the language. While there are two written languages,[1] neither is representative of any one dialect, and adopting a different dialect, especially Urban East Norwegian (spoken in and around Oslo), is considered inauthentic and looked down upon.

This section is structured as follows: I first introduce the classification that is typically applied to Norwegian dialectology (section 3.1). I then present the data that I work with (section 3.2). In section 3.3, I introduce previous approaches to automatic dialect classification, and in section 3.4, I present prior dialectometric work with Norwegian data. Next, I explain the classification approach (section 3.5). In section 3.6, I show and analyze the results, including general observations on the LIME scores, as well as an analysis of how the linguistic features most commonly used for dividing the Norwegian dialect landscape are (not) represented in the results, and finally other recurrent linguistic features in the results.

3.1 Norwegian dialects

The Norwegian dialect landscape is generally divided into four dialect groups: East Norwegian, West Norwegian, Trønder,[2] and North Norwegian (Jahr, 1990b, p. 10; Mæhlum and Røyneland, 2012, pp. 32–42; Barðdal et al., 1997, pp. 263–264; Hanssen, p. 118). Figure 3.1 shows the geographic areas in which the different dialect groups are spoken.

This division is based on linguistic properties that I later explain in subsection 3.6.2. To some degree, such linguistic boundaries also match certain natural borders. For instance, the linguistic border between the West and East Norwegian dialect groups

[1] I use Bokmål for the Norwegian examples in this chapter, as this is the written language used in the ScanDiaSyn corpus.

[2] Trønder dialects are mostly spoken in Trøndelag county in central Norway.

largely coincides with a mountain range separating the two geographic areas (Sandøy, 1991, p. 104).

The split into dialect groups is not entirely clear-cut, but complicated by several factors. There is ample variation within each group, and there exist dialects that act as transition zones between the more characteristic varieties of different dialect groups (Mæhlum and Røyneland, 2012, p. 29). Furthermore, individual regions within Northern Norway are influenced by linguistic contact in ways that do not apply to the majority of other Norwegian dialects (contact with Sámi languages and with dialects spoken by East Norwegian settlers) (Mæhlum and Røyneland, 2012, pp. 116; Jahr, 1990b, pp. 180, 182).

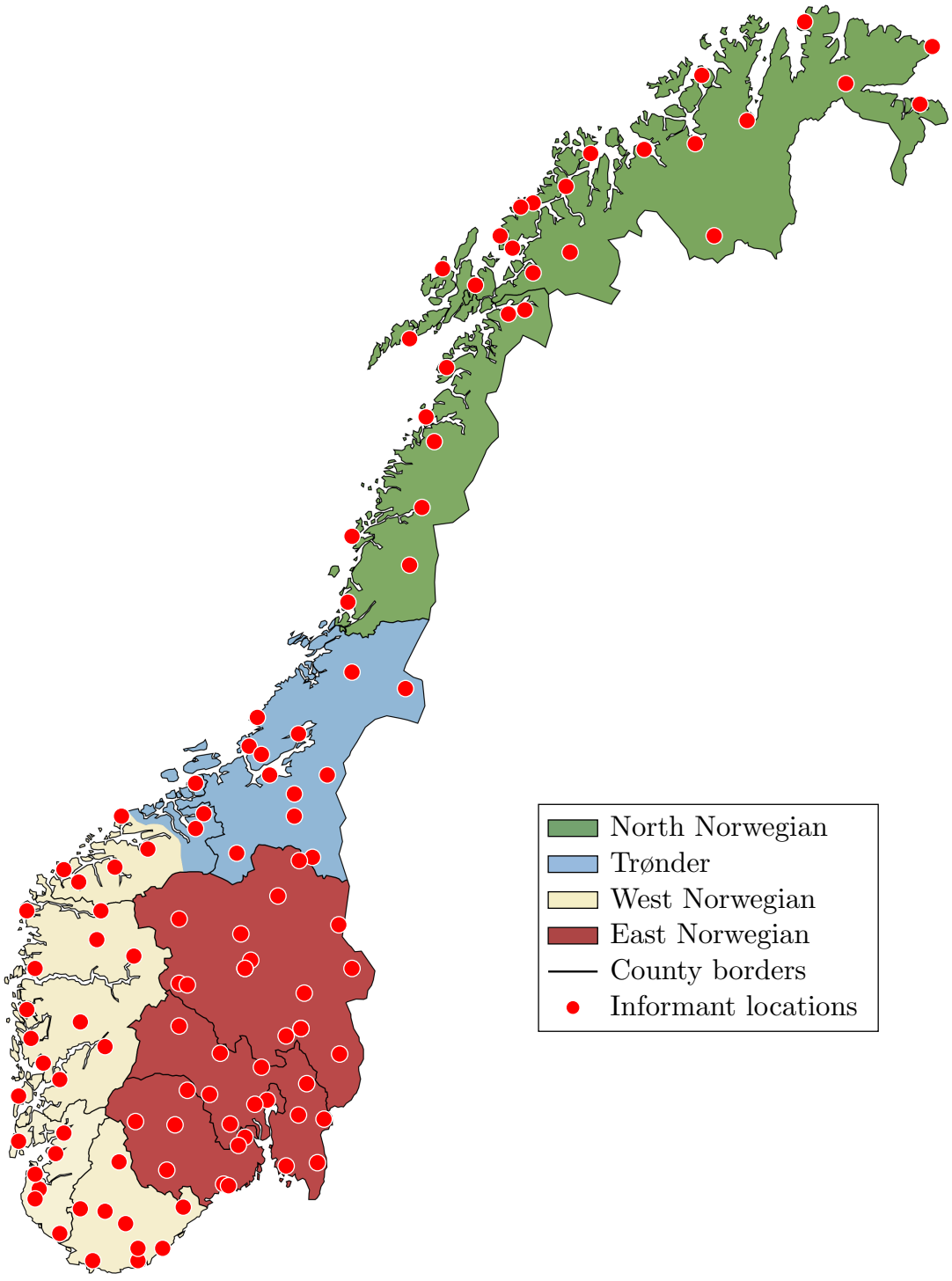


Figure 3.1: Dialect areas in Norway and ScanDiaSyn informant locations. The division into dialect areas follows the one by Mæhlum and Røyneland (2012, p. 178).

3.2 Data

I work with phonetically transcribed conversations from the Norwegian part of the ScanDiaSyn corpus (Johannessen et al., 2009).[3]

The data were obtained during interviews with informants or while recording conversations between informants. These interviews or conversations were transcribed both phonetically and in a written standard language. All utterances are instances of spontaneous speech.

I use utterances from all interviews/conversations that were transcribed both orthographically and phonetically. This includes over 116,000 utterances from 434 informants from 109 towns. Figure 3.1 shows where in Norway these towns are located. Approximately a quarter each of the informants consists of old women, young women, old men and young men.

East, West and North Norwegian are each represented by between 28 and 33 locations and between approximately 32,800 and 34,000 utterances. Only the Trønder dialect group is notably smaller: It is represented by around 15,900 utterances from fifteen locations.

Dialect group	**# of locations**	**# of informants**	**#, proportion of utterances**		**Mean utt. len.**
East Norwegian	33	131	32,802	28 %	13.7
West Norwegian	33	133	33,316	29 %	13.6
Trønder	15	65	15,903	14 %	12.7
North Norwegian	28	105	33,997	29 %	13.6
Total	109	434	116,018	100 %	13.5

Table 3.1: The class distribution in the ScanDiaSyn dataset. The mean utterance length is given in tokens per utterance.

3.2.1 Transcription

The interviews were transcribed twice, once in a very broad phonetic transcription that follows a custom transcription style and once in the written standard Bokmål.

Table 3.2 and Table 3.3 show how this custom transcription corresponds to IPA symbols for vowels and consonants, respectively. The tables are based on the ScanDiaSyn transcription manual (Johannessen et al., 2009, pp. 10–13) and further information on Norwegian phonology (Kristoffersen, 2000, pp. 13, 19–20, 22–25). These tables also contain the symbols I used for preprocessing the data. This is explained in more detail in subsection 3.2.2.

[3] Available at `http://tekstlab.uio.no/nota/scandiasyn/` under a CC BY-NC-SA 4.0 license.

IPA	ScanDiaSyn	Mine	IPA	ScanDiaSyn	Mine
/i/	i	i	/aʊ/	ao	ao
/y/	y	y	/æʉ/	æu	æu
/ʉ/, /ʊ/	u	u	/æɪ/	æi	æi
/u/	o	o	/ɛɪ/	ei	ei
/e/, /ɛ/, /ə/	e	e	/ɑɪ/	ai	ai
/ø/, /œ/, /ɵ/	ø	ø	/ɔɪ/	åi	åi
/o/, /ɔ/	å	å	/œʏ/	øy	øy
/æ/	æ	æ	/e.i/	e'i	**e.i**
/ɑ/, /a/	a	a			

Table 3.2: Norwegian vowels, as represented in the International Phonetic Alphabet, by the ScanDiaSyn project, and in this thesis. Where my transcriptions diverge from the ScanDiaSyn standard, entries are in boldface. (Here, this only applies to the notation of diphthongs). Other non-diphthong vowel sequences than the one in the last row are represented similarly.

To make it possible to directly align both transcriptions, both are carried out on a word level. The phonetic transcription does not show regular phonetic assimilation across word boundaries. Johannessen et al. (2009) refer to Papazian and Helleland (2005, p. 21) to argue that this makes it easier for humans to parse the transcription and that the phonological processes that occur across word boundaries in some dialects are so regular that seeing them occur *within* a word should be a clear sign for readers to predict that they also occur across word boundaries. *Irregular* assimilation across word boundaries is transcribed, however (p. 13).

Several letter sequences can either encode a single sound or a sequence of several sounds, e.g. ⟨rn⟩ for either /ɳ/ or /ʁn/, /ɾn/. This is intended for ease of transcription and reading (Johannessen et al., 2009, p. 11), although it entails the loss of information that is frequently used in descriptions of Norwegian dialects. Similarly, palatalization is not marked either.

Vowel length is only indicated in stressed syllables and monosyllabic words (Johannessen et al., 2009, pp. 11–12): the (first consonant of the) coda is doubled if the stressed syllable is short. Tonemes are not marked.

Each utterance is also transcribed into Bokmål. This is done on a word level; the syntax was not changed to match Bokmål syntax (Laake et al., 2011, p. 2). Bokmål allows some degree of freedom regarding word choice and several morphological details. The transcribers were free to use any of the valid lexical and morphological variants as they saw fit, but did not have to pick the ones closest to the dialect they were transcribing (Laake et al., 2011, p. 2). Correspondingly, there are some inconsistencies in the transcription. For instance, the question word /koʂʂn̩/ 'how' is sometimes

IPA	ScanDiaSyn	Mine	IPA	ScanDiaSyn	Mine
/p/	p	p	/f/	f	f
/b/	b	b	/s/	s	s
/t/, /c/	t	t	/s̩/	’s	**s̩**
/d/	d	d	/ʂ/, /ʃ/	sj	**ʂ**
/ʈ/, /ʁt/, /ɾt/	rt	rt	/ç/	kj	**ç**
/ɖ/, /ʁd/, /ɾd/	rd	rd	/h/	h	h
/k/	k	k	/t͡ʃ/	tj	**t͡ʃ**
/g/	g	g	/ɾ/, /ʁ/	r	r
/m/	m	m	/ɽ/	L	**ɽ**
/m̩/	’m	**m̩**	/ɽ̩/	’L	**ɽ̩**
/n/, /ɲ/	n	n	/l/, /ʎ/	l	l
/n̩/	’n	**n̩**	/l̩/	’l	**l̩**
/ɳ/, /ʁn/, /ɾn/	rn	rn	/ɭ/, /ʁl/, /ɾl/	rl	rl
/ŋ/	ng	**ŋ**	/ʊ/, /v/, /w/	v	v
/ŋ̩/	’ng	**ŋ̩**	/j/	j	j

Table 3.3: Norwegian consonants, as represented in the International Phonetic Alphabet, by the ScanDiaSyn project, and in this thesis. Where my transcriptions diverge from the ScanDiaSyn standard, entries are in boldface.

transcribed as the Bokmål word *hvordan* and sometimes as the synonymous term *åssen*.

The following excerpt from an interview recorded in an East Norwegian municipality (interview ID: *vang_ 02uk-sl.txt*) gives an impression of the different transcriptions:

(1) **Bokmål** når jeg har blitt eldre så prøver jeg lissom å
Phonetic når e ha vørrte elldre så prøve e lissåm å

holde mer på # d- dialekta mi enn hva jeg gjorde før
hallde mæir på # d- dialekto mi enn kå e joLe før

‘Having gotten older, I, like, try to insist more on using my dialect than I used to.’

3.2.2 Preprocessing

I normalize the data by removing all punctuation symbols, including characters like # that represent pauses within an utterance. I also remove all place names and names

of people. Additionally, I remove stuttering, aborted articulations and interjections (such as "mhm"). (The ScanDiaSyn documentation includes a list of dialect-neutral interjections (Johannessen et al., 2009, p. 30).) If an utterance contains less than three tokens after these steps have been applied, I skip the utterance.

Furthermore, I apply some changes to the phonetic transcription, as documented in Table 3.2 and Table 3.3. When possible, I replace symbol sequences that encode a single sound by a single (IPA) symbol. I also replace *L* with ɭ so that the data become case-independent.

This is how the previous utterance is encoded after these preprocessing steps:

(2)

Bokmål	når	jeg	har	blitt	eldre	så	prøver	jeg	lissom		
ScanDiaSyn	når	e	ha	vørrte	elldre	så	prøve	e	lissåm		
Mine	når	e	ha	vørrte	elldre	så	prøve	e	lissåm		
å	holde	mer	på	#	d-	dialekta	mi	enn	hva	jeg	gjorde
å	hallde	mæir	på	#	d-	dialekto	mi	enn	kå	e	joLe
å	hallde	mæir	på			dialekto	mi	enn	kå	e	joɭe
før											
før											
før											

'Having gotten older, I, like, try to insist more on using my dialect than I used to.'

In the rest of the chapter, all sounds and words between slashes are written in my adapted version of the ScanDiaSyn transcription style.

3.3 Automatic dialect disambiguation

A fair amount of research on *dialect disambiguation*—automatically discerning between different related dialects—has been made in recent years. Many of the results come from a range of tasks organized by the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) (Zampieri et al., 2017, 2018, 2019; Găman et al., 2020; Chakravarthi et al., 2021). Participants in these tasks have used many different machine learning techniques, including recurrent or convolutional neural networks, support vector machines (SVMs), BERT, naive Bayes classifiers and ensembles thereof.

In many (though not all) of these tasks, the winning systems encode the features as bags of character- and word-level n-grams and use SVMs as the classifier (e.g. the systems by Malmasi and Zampieri (2017), Bestgen (2017), Çöltekin et al. (2018) or Çöltekin (2020)). I base my dialect classification model on this; the details are described in section 3.5.

3.4 Norwegian dialectometry

While none of the VarDial tasks have focused on classifying Norwegian dialects, research in that area has been conducted.

Heeringa and Gooskens (2003) and Heeringa et al. (2009) cluster Norwegian dialects based on phonetic transcriptions and acoustic features and find that their results largely correspond to the findings of traditional dialectology and to speaker perceptions. Heeringa (2004, pp. 199–211) clusters Norwegian dialects into groups based on acoustic differences.

Gooskens and Heeringa (2006) investigate to what extent prosodic, phonetic and lexical distances between Norwegian dialects correlate with perceptual distances. Beijering et al. (2008) explore the correlation between phonetic distances and intelligibility ratings between Scandinavian dialects.

More recently, Kåsen et al. (2020) present a comparison of two different methods for quantifying dialect similarity, using a dataset that is similar to the dataset used in this thesis, in terms of how they were collected and the relatively coarse phonetic transcription style. They show that clustering based on edit distance works well for these data and produces results that agree with the traditional dialectology, and the same applies for clusters created using neural autoencoders if the training dataset is sufficiently large. Kåsen et al. conclude that "a coarse-grained transcription of speech is sufficient to replicate known dialectal boundaries."

3.5 Method

I represent each preprocessed utterance as a bag of n-grams: word-level uni- and bigrams, and character-level $\{1, 2, 3, 4, 5\}$-grams. All word-level n-grams are represented as a combination of their orthographic and phonetic representation. That is, the word-level unigrams corresponding to the beginning of Example 2 are: <SOS>når/når<EOS>, <SOS>jeg/e<EOS>, <SOS>har/ha<EOS>, and the corresponding word-level bigrams are <SOS>når/når<SEP>jeg/e<EOS>, <SOS>jeg/e<SEP>har/ha<EOS>, and so on. The meta-tokens <SOS>, <EOS>, and <SEP> stand for "start of sequence," "end of sequence," and "separator," respectively. I also use <SEP> to represent the word boundary in character-level n-grams. The word *når* 'when' for instance consists thus of the character bigrams <SEP>n, nå, år, and r<SEP>.

These n-grams are numerically encoded using TF-IDF (term frequency, inverse document frequency) weighting. Only the top 5000 features (in the training data) are considered in the TF-IDF encoding step, features that appear more rarely are ignored when training and testing the model. This encoding is done using the scikit-learn library for Python (Pedregosa et al., 2011).

The classifier is a support vector machine (SVM) with a linear kernel, also as implemented in scikit-learn. The four-way classification is performed by training one one-versus-rest classifier per dialect group.

Each of these classifiers produces a prediction for a given input instance. The confidence score for a classifier's prediction is proportional to the distance between the

input instance's representation in vector space and the classifier's decision hyperplane. The prediction probability distribution that LIME works with, $f(x)$, is the result of applying the softmax function to the classifiers' confidence scores.

3.6 Results

The results section is structured as follows: I first present the performance of the model and general information on the LIME-based importance scores in subsection 3.6.1. I then focus on the top 50 features per dialect group and analyze to what extent they reflect the linguistic features that are traditionally considered the most important distinctive features in Norwegian dialectology (subsection 3.6.2) and other linguistics features (subsection 3.6.3).

3.6.1 General observations

I train and test the dialect classification model in ten initializations, each on a different train-test split of the dataset, and extract LIME importance scores from each of these runs. All of the scores in this section are mean values across all ten runs. The average model accuracy is 78.6 % and the average (macro-averaged) F_1 score is 77.1 %.

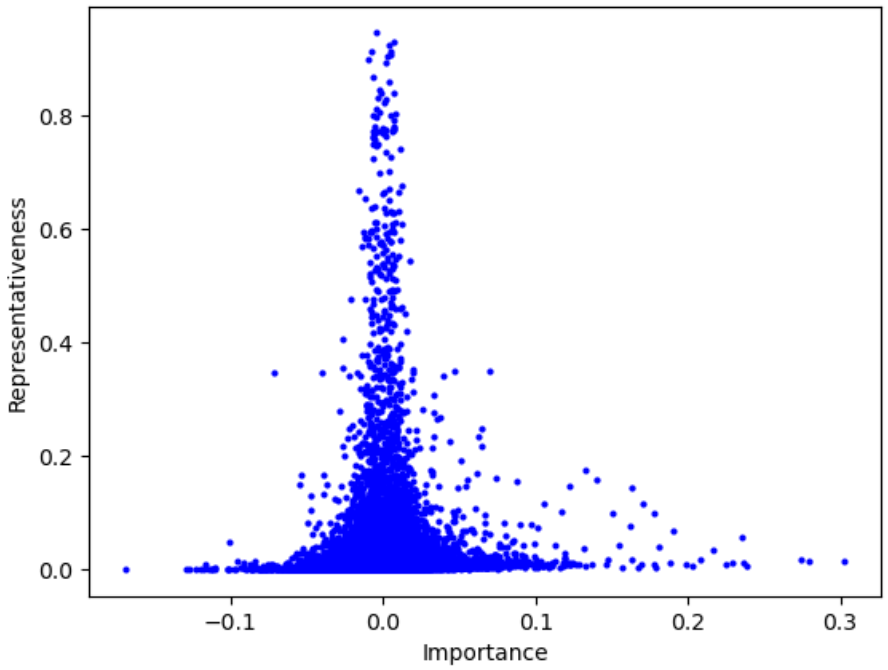


Figure 3.2: Representativeness values by LIME importance score per feature-label combination.

Importance values range between -0.17 and +0.30, with no significant distribution differences for the different dialect groups. There is only a marginal correlation

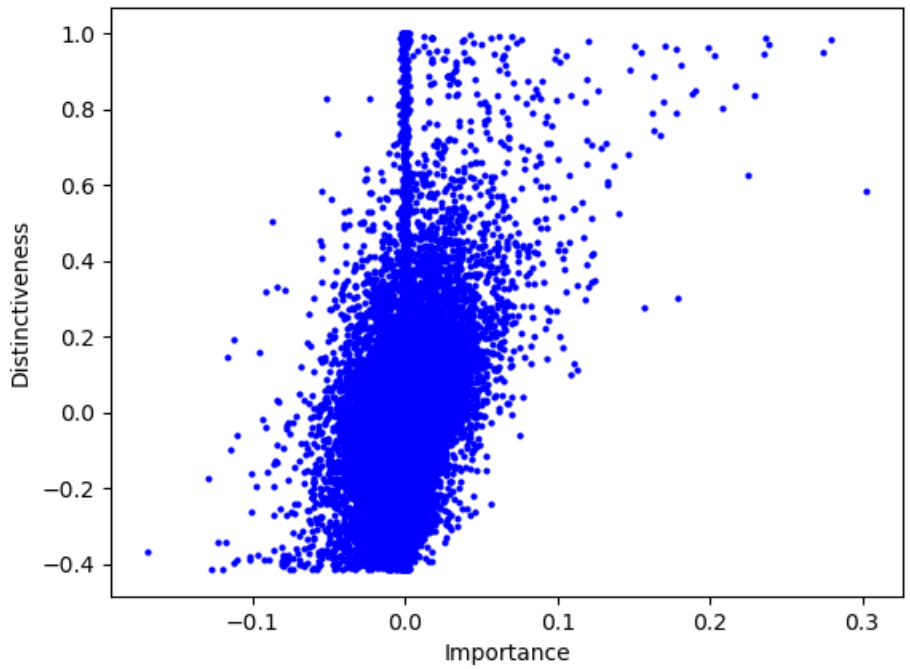


Figure 3.3: Distinctiveness values by LIME importance score per feature-label combination.

between a feature's importance score for a label and the corresponding representativeness value, i.e. in which proportion of the instances with that label it is present. This is not very surprising, as most features with high representativeness scores are representative of *all* dialect groups. This is for instance the case with almost all character unigrams. The correlation coefficient (Pearson's R) between importance and representativeness is between 0.03 and 0.05 for the different dialect groups, and this is also illustrated in Figure 3.2. However, the importance score does correlate with the distinctiveness score, that is, features with higher importance scores for a label tend to mostly occur in utterances with that gold-standard label (Figure 3.3). The correlation coefficient between importance and distinctiveness is between 0.37 and 0.41 for the different dialect groups.

Table 3.4 shows the importance scores for each label for a sample sentence, the West Norwegian utterance *ja da [.] vi har [—] de siste årene nå så har vi vært heldige* /ja då me he di sisste åran nå så he mi værrt helldi/ "Yes. We have—the past few years now, we've been lucky." (Note that these are the LIME scores for this specific utterance and not the global importance scores.) This utterance is represented by 100 features, the vast majority of which have LIME scores that are close to zero. This utterance was correctly predicted as West Norwegian and that prediction is also reflected by the distribution of importance scores: in this case, only the West Norwegian label is associated with (more than marginally) positive importance scores, while the importance scores for the other combinations of features and labels tend to be close to zero (signifying that a feature is insignificant for predicting the given label) or negative (indicating that the presence of the feature lowers the likelihood

Word	**West** (actual & predicted label)		**East**		**Trønder**	**North**	
ja /ja/ "yes"							
da /då/	0.17	då<SEP>	-0.07	då<SEP>	-0.07	då<SEP>	
"then"	-0.06	<SEP>då					
vi /me/ "we"	0.11	<SOS>vi/me<EOS>				-0.08	<SOS>vi/me<EOS>
har /he/	0.08	he<SEP>	-0.06	he<SEP>			
"have.PRES"	0.05	<SOS>har/he<EOS>					
de /di/	0.06	<SEP>di					
"the.PL"							
siste /sisste/ "last.DEF"							
årene /åran/ "years.DEF"							
nå /nå/	-0.11	<SEP>nå					
"now"	0.08	nå<SEP>					
så /så/ "so"							
har /he/	0.05	<SOS>har/he<EOS>					
"have.PRES"	0.08	he<SEP>					
vi /mi/ "we"	0.13	<SOS>vi/mi<EOS>	-0.06	<SOS>vi/mi<EOS>		-0.06	<SOS>vi/mi<EOS>
	0.10	<SEP>mi	-0.09	<SEP>mi			
vært /værrt/ "been"							
heldige /helldi/ "lucky.PL"							

Table 3.4: LIME scores for a (correctly predicted) West Norwegian utterance. Features with importance scores between -0.05 and +0.05 are omitted to preserve space. No word bigrams have importance scores that lie below/above that threshold.

of the given label being predicted). It should be noted that importance scores can seem contradictory: in the sample sentence, the words *da* /då/ and *nå* /nå/ are represented by a feature <SEP>då (<SEP>nå) "/då/ (/nå/) is a prefix (or full word)" and another feature då<SEP> (då<SEP>) "/då/ (/nå/) is a suffix (or full word)," where the former receives a negative importance score and the latter a positive one.

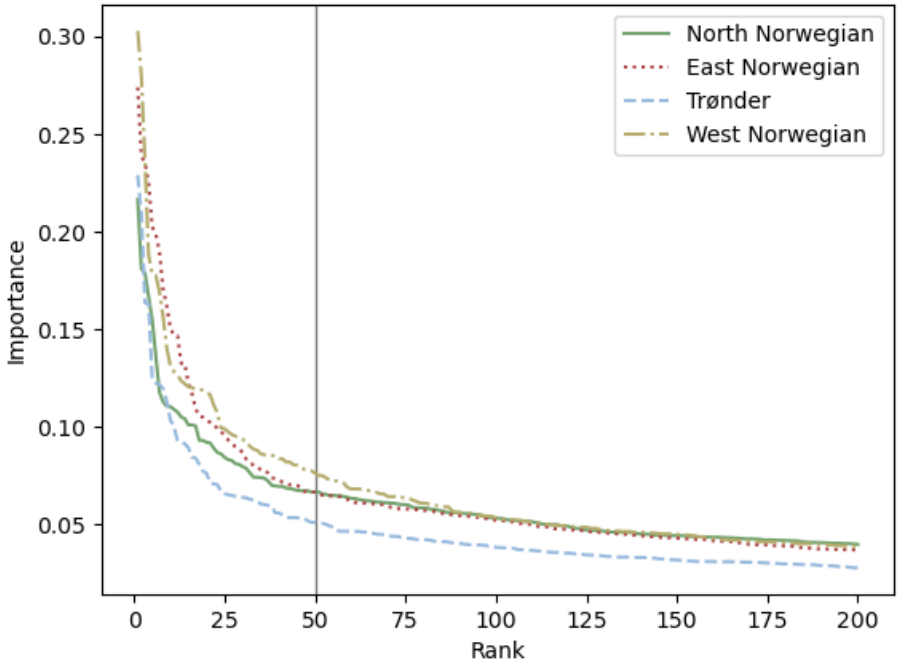


Figure 3.4: Importance score by rank and dialect group.

In the following two sections, I qualitatively examine the 50 features with the highest importance scores per predicted label. I chose this threshold to strike a balance between only analyzing features with relatively high importance scores and having a sizable selection of features to analyze. Figure 3.4 shows the importance scores for the features included in the analysis as well as the scores of the succeeding ranks. The selected features show a correlation between importance score and distinctiveness, as illustrated in Figure 3.5. The following analysis includes the 50 features per class that have the highest importance scores.[4] These features tend to mostly include variants of high-frequency words, as well as some common short sequences of phonemes. Only one of those high-importance features is clearly about a conversation topic (rather than lexical choice or pronunciation): the trigram ami in North Norwegian utterances, which usually appears in the word *samisk* 'Sámi' and inflected versions thereof. This is not a surprise as most Sámi cultural centres are located in the Northern part of the country.

[4]I share tables with the 200 most important features per dialect group at https://github.com/verenablaschke/ma-thesis/tree/main/models/dialects.

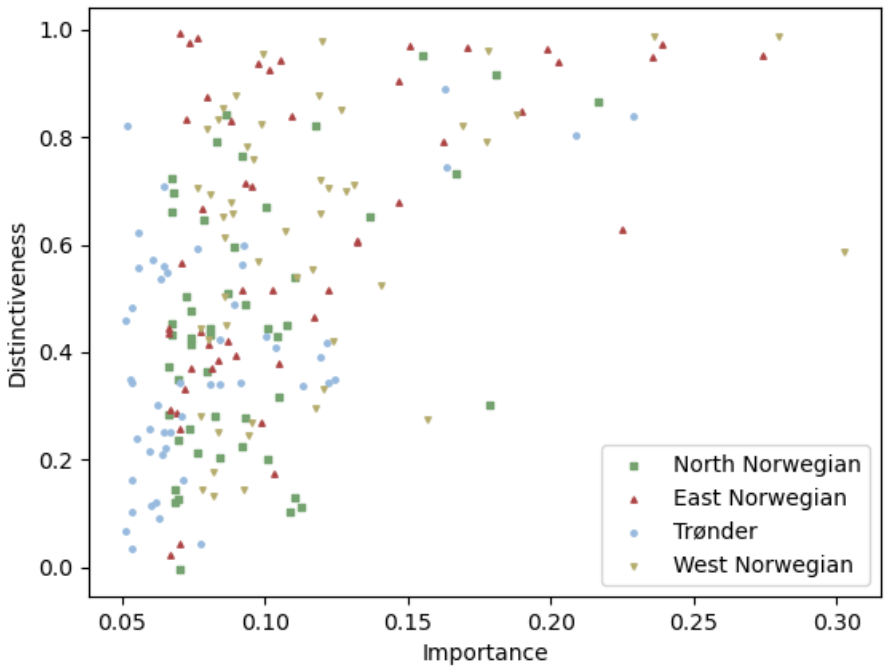


Figure 3.5: Distinctiveness by importance score for the 50 highest-ranking features per label.

3.6.2 Major linguistic features

In this section, I introduce the linguistic characteristics that are typically discussed in Norwegian dialectology and point out which of these can be found in the ScanDiaSyn data and whether they are considered as important by LIME.

Different dialectologists use somewhat different sets of linguistic features to characterize the different dialect groups. In this section, I present a summary of those features that are regularly brought up in the literature on Norwegian dialectology.

Sandøy (1991, pp. 113–115) uses the features detailed in the first of the following passages ("Infinitive endings...") to discern between twelve dialect groups (that are subgroups of the four groups I use in this thesis). Mæhlum and Røyneland (2012, pp. 32–42) base their classification on the features in all of three of the following passages. These are also the features considered most important by Kåsen et al. (2020).

Infinitive endings and endings of feminine nouns

One prominently discussed group of features is concerned with the different ways in which word-final vowels of infinitives or certain feminine nouns have changed. The following explanation summarizes the overviews by Mæhlum and Røyneland (2012, pp. 33–35) and Sandøy (1991, pp. 113–114).

In Old West Norse, infinitive forms of verbs with more than one syllable ended in /-a/, as did the so-called 'weak' feminine nouns (that is, feminine nouns ending in a vowel sound rather than a consonant). The following types of dialects have emerged with regard to how this ending has (or has not) changed:

- *A-mål* 'a-speech': In these dialects, all such words still end in /-a/ (or another non-schwa vowel).
- *E-mål* 'e-speech': The endings of both infinitives and weak feminine nouns were reduced to a schwa.
- Apocope: Both infinitives and weak feminine nouns have undergone apocope.
- *E/a-mål* 'e/a-speech': Only the infinitive endings were reduced to a schwa, weak feminine nouns still end in /-a/ (or another non-schwa vowel).
- *Jamvektsmål* 'balance-speech': Whether or not the final vowel was reduced or not depends on the length of the root of the word. Only infinitives and weak feminine nouns with short roots retained endings with full endings, whereas words with roots whose rhyme contained a long vowel and/or multiple consonants now end in /-ə/.
- *Jamvekt* with apocope: These dialects behave like the previous group, but the final vowel of a word with a long root was dropped.

For classifying to which of the major dialect groups a doculect belongs, *jamvekt* and apocope are often considered the most distinctive indicators (Mæhlum and Røyneland, 2012, pp. 32–42). East Norwegian dialects fall into the *jamvekt* group (with /-ə/) (Mæhlum and Røyneland, 2012, p. 46). Trønder dialects exhibit *jamvekt* with apocope (p. 76), and West Norwegian dialects are instances of *a-mål* and *e-mål* (p. 90). The different North Norwegian dialects fall into all of the listed groups except for either of the *jamvekt* types (pp. 106–107).

All of the phenomena listed in this section can be found in the data, albeit not overtly encoded. However, they can at least be partially found when inspecting common infinitive forms in the data: The by far most common (multisyllabic) infinitives in the ScanDiaSyn data are *(å) være* '(to) be,' *(å) gjøre* '(to) do,' and *(å) komme* '(to) come.' All three verbs are in the group of verbs whose ending is *not* reduced in *jamvekt* dialects (cf. Hanssen, 2010, p. 84); therefore knowing the infinitive forms of these verbs for a given dialect is *not* sufficient for figuring out exactly which suffix group the dialect belongs to, although it can be used to narrow down the options, as shown in Table 3.5.

Versions of *være* and *gjøre* are represented among the features with high importance scores for instances predicted as East Norwegian or Trønder: `ære`<SEP> in East Norwegian and `rrå` and `rra`<SEP> in Trønder; all indicating full vowel endings, as expected. All of these features represent both *være* and *gjøre* at the same time (and in one case, the verb *(å) fare* '(to) drive' as well). None of the highest-ranking features include versions of *komme* despite it also appearing frequently in the data (but there are also no other frequent verbs in the dataset whose stem ends in *-mm*).

No feminine nouns (or features that clearly encode the ending of a feminine noun) are among any dialect group's top 50 labels. However, even the most frequently

Type	**(å) være** ‘(to) be’	**(å) gjøre** ‘(to) do’	**(å) komme** ‘(to) come’
A-mål, jamvekt	væra, vårrå, værra	jørra, jøra, jera	kåmma, kåmmå
E-mål, e/a-mål	være	jøre, jære	kåmme, kåme
Apocope	vær, væ	jør, jær, jørr	kåmm

Table 3.5: The most common infinitive forms of the three most common (non-monosyllabic) verbs in the ScanDiaSyn corpus, grouped by the type of ending. The examples in this table are by no means exhaustive. The *jamvekt* subgroup here includes both dialects with and without apocope.

Group	**Feature**	**Imp.**	**Rep.**	**Dist.**	**Context (bokmål/pron.)**
East	`æra<SEP>`	0.08	0.01	0.44	være/væra(0.7) ‘(to) be’ gjøre/jæra(0.1) ‘(to) do’
Trø.	`rrå`	0.09	0.02	0.56	være/vårrå(0.7) ‘(to) be’ fare/fårrå(0.1) ‘(to) drive’
	`rra<SEP>`	0.05	0.02	0.24	være/værra(0.2) ‘(to) be’ gjøre/jørra(0.2) ‘(to) do’

Table 3.6: Features encoding infinitive endings that are among the top 50 most important features per dialect group. The context column contains the most common token-level context for character n-grams (numbers in parentheses indicate the proportion of context tokens a character n-gram comes from).

appearing weak feminine nouns (*klasse* 'class,' *uke* 'week,' and *hytte* 'hut') occur significantly less often than the most common verbs.

Prosody

A second distinctive feature is the realization of the two tonemes that exist within the context of Norwegian pitch accent. When a word has accent 1, speakers of East Norwegian and Trønder begin with a low pitch whereas West and North Norwegian dialect speakers tend to begin with a high pitch (Mæhlum and Røyneland, 2012, p. 37). An experiment by Gooskens (2005) shows that intonation information plays a significant role when Norwegians are asked to determine where a dialect speaker is from. This is also confirmed by van Ommeren and Kveen (2019). Toneme information is *not* encoded in the ScanDiaSyn dataset.

However, Mæhlum and Røyneland (2012, pp. 36–37) mention another prosodic feature that correlates with the toneme patterns: word-level stress in particle verbs and in many Greek and Romance loanwords. Generally, the last syllable of such a loanword (and the particle in a particle verb) are stressed in North and West Norwegian, whereas the first syllable (and the verb) are stressed in East Norwegian and Trønder (Hanssen, 2010, pp. 58–59; Mæhlum and Røyneland, 2012, pp. 37, 78). While the stress pattern in particle verbs is not always overtly represented in ScanDiaSyn,[5] it is encoded in some loanword transcriptions, such as pronunciations of *spesiel* 'special,' which is transcribed as either `spessiel` (with stress on the first syllable) or `spesiell` (with stress on the second syllable). None of the top 50 features encode stress information. The highest-ranking feature to do so is `ssi` in Trønder (rank 59 with a mean importance score of 0.05), which most often appears in phonetic transcriptions of the words *spesielt* 'special, especially' and *musikk* 'music.'

Retroflex flap

Another important feature is the presence or absence of the retroflex flap. In many dialects, the Old Norse phoneme /l/ changed to /ɽ/ in many phonological environments, and often, Old Norse /rð/ also changed to /ɽ/ (instead of /r/) (Sandøy, 1991, p. 185).

These changes are characteristic of East Norwegian and Trønder dialects, whereas West Norwegian dialects do not have this consonant, and the North Norwegian dialect area contains dialects with and without /ɽ/ (Mæhlum and Røyneland, 2012, pp. 36, 184).

The East Norwegian and Trønder dialects contain several high-ranking features that include /ɽ/ (Table 3.7). In most cases, these features are character-level n-grams that only appear in one or a few words in the corpus at large, although these words tend to be quite common. However, the unigram ɽ achieves a relatively high ranking among the East Norwegian features (despite not making it past the rank threshold): it is at rank 56 with an importance score of 0.06.

[5] It is only transcribed when the stress lies on the verb and the stressed syllable within the verb contains a short vowel.

Group	Feature	Imp.	Rep.	Dist.	Context (bokmål/pron.)
East	øɽɽ	0.15	0.01	0.68	folk/føɽɽk(0.2) ‘people’
	<SEP>bɽe	0.11	0.01	0.84	ble/bɽe(0.8) ‘became’
	ɽæi	0.08	0.01	0.87	blei/bɽæi(0.7) ‘became’
	<SEP>oɽ	0.07	0.00	0.56	ord/oɽ(0.9) ‘word’
Trø.	eɽ<SEP>	0.06	0.02	0.62	vel/veɽ(0.6) ‘well’
	<SEP>væɽ	0.06	0.02	0.54	vel/væɽ(1.0) ‘well’
	<SEP>veɽ	0.06	0.01	0.71	vel/veɽ(1.0) ‘well’
	øɽ	0.06	0.03	0.25	sjøl/ʂøɽ(0.4) ‘self’
	æɽ<SEP>	0.05	0.02	0.48	vel/væɽ(0.7) ‘well’

Table 3.7: Features with high importance values that contain /ɽ/.

3.6.3 Other linguistic features

The previously mentioned features are by far not the only features included in classifications and descriptions of Norwegian dialects. This section presents some of the other linguistic features with high importance scores that are often discussed in Norwegian dialectology, despite not always being considered the most essential for deciding where the borders between the dialect areas should be drawn.

Personal pronouns

There is also ample variation in the variants of personal pronouns. The first person singular pronoun *jeg* is pronounced /e(g)/ or /æg/ in large parts of the country, but with an initial /j-/ (/je/ or /jæ(i)/) in much of the East Norwegian area (Jahr, 1990b, pp. 22–23). In parts of West Norway and Trøndelag, the variant /i/ is also in use (Jahr, 1990b, pp. 22–23). Apart from the geographic distribution of /i/, the literature generally does not show such a subdivision and tends to lump together the forms without /j-/ in the remaining regions. Table 3.8 shows the first personal singular forms that rank among each dialect group’s 50 highest-scoring LIME features. These results clearly show the presence of an initial glide in East Norwegian 1.SG forms. The North Norwegian form with the highest importance score is /æ/. While it is not the only form used in that dialect group, Jahr and Skare (1996, p. 36) already remarked upon its spreading popularity several decades ago. Additionally, the results highlight several West Norwegian pronoun variants: <SOS>jeg/i<EOS>, <SOS>jeg/ei<EOS> and eg<SEP>. While these are generally not used to characterize the entire West Nowegian dialect group, they are characteristic for several dialects within that group Sandøy (1990, pp. 71, 74, 76, 79). Despite being a lot less commonly discussed by dialectologists than the first person singular nominative pronoun, two versions of the accusative form also receive high importance scores: Trønder /mæ/

Pron.	Group	Feature	Imp.	Rep.	Dist.	Context (bokmål/pron.)
1.SG NOM	North	<SOS>jeg/æ<EOS>	0.07	0.16	0.43	
	East	<SOS>jeg/je<EOS>	0.19	0.07	0.85	
		<SOS>jeg/jæ<EOS>	0.17	0.12	0.97	
		<SEP>jæi	0.15	0.02	0.91	jeg/jæi(1.0)
		jæ<SEP>	0.11	0.12	0.94	jeg/jæ(1.0)
	West	<SOS>jeg/i<EOS>	0.30	0.02	0.59	
		<SOS>jeg/ei<EOS>	0.17	0.01	0.82	
		eg<SEP>	0.09	0.16	0.68	jeg/eg(0.9)
1.SG ACC	Trø.	<SOS>meg/mæ<EOS>	0.06	0.02	0.11	
	East	mæi<SEP>	0.08	0.01	0.67	meg/mæi(0.9)
1.PL NOM	Trø.	<SOS>vi/åss<EOS>	0.10	0.02	0.43	
	West	<SOS>vi/mi<EOS>	0.12	0.02	0.66	
		<SOS>vi/me<EOS>	0.10	0.08	0.57	

Table 3.8: First person pronouns in the top 50 most important features per dialect group. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each character n-gram appears (along with the relative frequency of this word being the origin).

and East Norwegian /mæi/.

Vi, the first person plural pronoun, is replaced by /me/ or /mi/ in many West Norwegian and some East Norwegian dialects, and by /åss/ in some other East, West and Trønder Norwegian dialects (Mæhlum and Røyneland, 2012, p. 183). Table 3.8 also shows the first person plural forms that are among the most important features, as determined by LIME. These clearly reflect the West Norwegian tendency to use /me/ or /mi/. The results also include Trønder /åss/, which—while also attested in other dialect groups—is more characteristic of Trønder in the ScanDiaSyn data (about half of the occurrences of <SOS>vi/åss<EOS> appear in just 14 % of the data).

Pron.	Group	Feature	Imp.	Rep.	Dist.	Context (bokmål/pron.)
2.SG NOM	East	<SOS>vet/vett<SEP>du/du<EOS>	0.08	0.01	0.38	
		<SEP>ru	0.08	0.03	0.41	du/ru(0.8)
	Trø.	<SOS>vet/vet<SEP>du/du<EOS>	0.09	0.03	0.34	
		<SOS>du/u<EOS>	0.06	0.01	0.22	
	West	do<SEP>	0.13	0.01	0.70	du/do(0.9)
2.SG ACC	Trø.	<SOS>deg/dæ<EOS>	0.06	0.01	0.12	
2.PL	North	dåkke	0.07	0.01	0.50	dere/dåkker(0.7)

Table 3.9: Second person pronouns in the top 50 most important features per dialect group. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each character n-gram appears (along with the relative frequency of this word being the origin).

The second person singular (nominative) pronoun is not commonly presented as a particularly important feature for distinguishing between dialect groups. In most dialects, it is /du/, although (when unstressed) it is reduced to /ru/ in some East Norwegian dialects (Hårstad and Opsahl, 2013, p. 88; Endresen, 1990, p. 97; Lie, 1990, p. 184). The high-ranking features include East Norwegian /ru/ as well as a West Norwegian form /do/. The latter does in fact most commonly appear in West Norway in the ScanDiaSyn data (although /du/ is nevertheless the most frequent pronunciation in that part of the country) but it is not remarked upon in the descriptions of West Norwegian dialects by Sandøy (1990), Hanssen (2010, pp. 168, 176, 185) or Mæhlum and Røyneland (2012, p. 91). Two word bigrams with different variations of *vet du* 'you know; do you know' also have high importance scores among the East Norwegian and Trønder features, but both include the common form /du/ and only differ in the vowel length of /vet(t)/. Trønder also has the high-importance variant /u/, which is not commonly pointed out in descriptions of the dialect group. The

accusative form *deg* usually also goes unremarked in dialectologist literature, but the Trønder pronunciation /dæ/ has a relatively high importance score (similarly to the Trønder 1.SG.ACC form /mæ/).

Different dialects use different lexemes for second person plural pronouns. In East and West Norwegian areas, variants of /di, de/ (NOM) and /dere/ or /dVkk, dVkkV(r/n)/ (ACC or regardless of case) prevail (Papazian, 2008, pp. 80–86). Speakers of Trønder dialects use /di, de/ (NOM) and /dåkk/ (ACC or regardless of case) (Papazian, 2008, p. 86), whereas North Norwegian dialects do not make any case distinction and use forms resembling /dåkk(er)/ (Papazian, 2008, p. 87). Of these forms, only the North Norwegian /dåkker/ appears in the top 50 features per dialect group (see Table 3.9).

Pron.	Group	Feature	Imp.	Rep.	Spec.	Context (bokmål/pron.)
3.SG	North	<SOS>hun/o<EOS>	0.09	0.01	0.28	
		ho<SEP>	0.08	0.04	0.21	hun/ho(0.9)
3.PL	East	<SEP>ræi	0.10	0.01	0.17	de/ræi(0.4)
		dømm<SEP>	0.07	0.02	0.97	de/dømm(0.8)
		ømm<SEP>	0.07	0.02	0.83	de/dømm(0.6)
	Trø.	<SEP>æmm	0.16	0.02	0.74	de/æmm(0.9)
	West	<SOS>de/dei<EOS>	0.09	0.01	0.85	
		<SOS>de/dæi<EOS>	0.08	0.06	0.69	
		<SEP>dei	0.08	0.01	0.70	de/dei(0.9)

Table 3.10: Third person pronouns in the top 50 most important features per dialect group. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each character n-gram appears (along with the relative frequency of this word being the origin).

The most common variant of the third person feminine singular pronoun is /ho/, although /hu(n)/ is also common in East Norwegian and /hon/ in parts of West Norway (Hanssen, 2010, p. 110). Only the North Norwegian dialect group contains features encoding this pronoun in its top 50 list, and in this case this is the prevailing form /ho/ as well as a reduced variant /o/.

There are two common variants of the 3.PL pronoun: /di, de(i)/ and /dVmm/. West Norwegian dialects use the former variant, Trønder dialects the latter (/dæmm/ or /dåmm/), and both are found in East Norway (/demm, domm, dømm/, /di/) and North Norway (/di/, /dæmm/) (Mæhlum and Røyneland, 2012, pp. 52, 78, 91, 109). Table 3.10 shows the third person pronouns that are among each dialect group's highest-ranking 50 LIME features. The East Norwegian dialect group includes /dømm/ as well as a form with the /d-/–/r-/ correspondence that is also present in

the 2.SG pronouns. As expected from the literature, the top LIME features for the West Norwegian dialects contain forms without -m (/dei, dæi/). The high ranking Trønder form is /æmm/, which resembles but is not identical to the form /dæmm/ that is expected from the literature. The dropped initial /d-/ is also repeated in the previously mentioned Trønder second person singular pronoun form /u/.

Negation

Group	Feature	Imp.	Rep.	Dist.	Context (bokmål/phon.)
North	<SEP>ikk	0.08	0.08	0.36	ikke/ikke(0.9)
East	<SEP>tte	0.24	0.01	0.97	ikke/tte(1.0)
	<SOS>ikke/itte<EOS>	0.24	0.06	0.95	
	çi<SEP>	0.22	0.01	0.63	ikke/çi(0.6)
Trø.	<SOS>ikke/itt<EOS>	0.16	0.14	0.89	
	<SEP>itt	0.12	0.15	0.42	ikke/itt(1.0)
	<SEP>tt	0.08	0.00	0.04	ikke/tt(1.0)
West	<SOS>ikke/ʂe<EOS>	0.24	0.01	0.99	
	<SEP>ççe	0.12	0.00	0.55	ikke/ççe(1.0)
	iʂʂ	0.11	0.01	0.63	ikke/iʂ̪ʂe(0.9)
	<SEP>t͡ʃe	0.10	0.05	0.96	ikke/t͡ʃe(0.9)
	ʂ̪ʂe<SEP>	0.09	0.01	0.61	ikke/iʂ̪ʂe(0.8)

Table 3.11: Variants of the negation *ikke* in the top 50 most important features per dialect group. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each character n-gram appears (along with the relative frequency of this word being the origin).

The negation word *ikke* is pronounced in many different ways across the country. The most common variant is /iççe/, but /itt/[6] is characteristic of Trønder, /itte/ is used in many East Norwegian dialects, and /ikke/ is used in some parts of North and East Norway (Jahr, 1990b, pp. 20–21). In West Norway, /iççe/ also appears alongside /itt͡ʃe/ and (in the city of Bergen) /iʃʃe/ (Mæhlum and Røyneland, 2012, pp. 91, 50). All of this is partially reflected in the results (Table 3.11). As in the literature, /(i)tte/[7] is the most commonly used form in East Norway, although /(iç)çi/ also

[6]Technically, the palatalized version is typical for Trønder (/ɪcc/ in IPA), but the ScanDiaSyn transcription system does not differentiate between palatal and alveolar stops.

[7]In spoken Norwegian, the initial /i-/ in *ikke* is often dropped.

ranks high. In the West Norwegian group, /ççe/, /(it)t͡ʃe/ and especially the Bergen variant /iʃʃe/ have high importance scores. The North Norwegian group only has one high-ranking feature representing the negation: /ikk(e)/. In the Trønder area, several n-gram representations of /(i)tt/ are ranked as important, as expected from the literature.

Question words

Most Norwegian question words begin with *(h)v-*. This initial sound is realized as /k-/ or /kv-/ in most dialects, with the exception of some of the dialects spoken in East or North Norway, where it is instead pronounced /v-/ (Sandøy, 1991, pp. 79–80). Two question words make it into the top 50 lists, namely to variants of *hva* 'what': East Norwegian <SOS>hva/va<EOS> and West Norwegian kå<SEP> (which is most commonly a subtoken of the <SOS>hva/kå<EOS>). While these represent typical variants of some East or West Norwegian question words, this brief list is very far from exhaustive when it comes to the full set of question words and local variations thereof.

Lexical variation

Works on Norwegian dialectology tend to briefly reference lexical variation but not go into detail (cf. Sandøy (1991, p. 104) and Hanssen (2010, pp. 114–115)). Gooskens and Heeringa (2006) find that lexical variation correlates significantly less strongly with dialect speakers' perceptual distances than differences in pronunciation (albeit with the caveat that their methodology might not encourage naturalistic lexical variation).

Group	Feature	Imp.	Rep.	Dist.	Context (bokmål/phon.)
East	<SEP>çue	0.07	0.00	0.33	tjue/çue(0.9) 'twenty'
	ræd	0.07	0.00	0.02	tretti/træddve(0.5) 'thirty'
North	yv	0.10	0.01	0.44	tjue/tyve(0.4) 'twenty' syv/syv(0.3) 'seven'

Table 3.12: Numerals in the top 50 most important features per dialect group. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word(s) in which each character n-gram appears (along with the relative frequency of this word being the origin).

In Norwegian, some numerals have both older and more recently introduced forms that exist in parallel: *syv* and *sju* 'seven,' *tyve* and *tjue* 'twenty,' and *tredve* and *tretti* 'thirty.'[8] Kvale and Foldvik (1997) found that there are some geographic patterns as to which forms are used: *sju* is especially common in the North and Trøndelag (grouped together in that article), and *tjue* is especially common in West Norway.

[8]The forms *tyve* and *tredve* are currently not part of written Bokmål, but I use them here to differentiate between the different lexical forms without having to specify phonetic details.

According to the authors, there are smaller differences in the usage of word forms for 'thirty,' although *tretti* is most common in the West. The top 50 lists contain three features that correspond to numerals (Table 3.12), none of which match Kvale and Foldvik's observations very closely. These features are the East Norwegian *tjue* and *tredve*, as well as the North Norwegian bigram `yv` that usually appears in *syv* and *tyve*. Unlike East Norwegian *tredve* which only appears slightly more often in that dialect group than you would expect if the occurrences were randomly distributed (30 % of the occurrences appear in a group that constitutes 28 % of the data), the other two features have fairly high specificity scores, indicating that the usage pattern of numerals may have changed in the past few decades or that the ScanDiaSyn data and Kvale and Foldvik's data contain different patterns for other reasons.

Noe(n) and *mye*

	Group	Feature	Imp.	Dist.	Rec.	Context (bokmål/pron.)
noe(n)	West	`nåkke`	0.12	0.01	0.71	noe/nåkke(0.6) noen/nåkken(0.3)
	Trø.	`<SOS>noe/nå<EOS>`	0.08	0.05	0.42	
	East	`nok`	0.10	0.01	0.52	noe/nokko(0.4)
		`<SOS>noe/no<EOS>`	0.09	0.04	0.51	
	North	`<SOS>noe/nåkka<EOS>`	0.09	0.02	0.84	
		`<SEP>nån`	0.07	0.02	0.48	noen/nån(0.7)
mye	Trø.	`myt`	0.08	0.01	0.34	mye/mytti(0.7)
		`<SEP>myt`	0.08	0.01	0.34	mye/mytti(0.7)
		`my<SEP>`	0.06	0.02	0.57	mye/my(1.0)
	East	`çy`	0.09	0.01	0.39	mye/myççy(0.4)

Table 3.13: Variants of *noe(n)* 'some, someone, something' and *mye* 'much' in the top 50 most important features per dialect group. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each character n-gram appears (along with the relative frequency of this word being the origin).

Several features with high importance scores represent variants of the words *noe(n)* 'some, something, someone' and *mye* 'much.' Both are high-frequency words that come in two main versions: with and without a /k/ (or other consonant) in the middle.

This variation is even represented in the two different orthographies (compare Bokmål *noe(n)* and *mye* and Nynorsk *noko/nokon/nokre* and *mykje*), but it is not commonly remarked upon in traditional Norwegian literature as an identifying trait for any of the dialect groups (see for instance the summaries of important identifying traits by dialect group by Hanssen (2010, pp. 125, 155, 163–164, 187–188) and Mæhlum and Røyneland (2012, pp. 45–54, 76–80, 89–94, 106–111)). The LIME results also do not show a clear separation here: East and North Norwegian both have high-ranking versions of *noe(n)* with and without /k/, but West Norwegian and Trønder both have only one high-ranking variant: /nåkke/ and /nå/, respectively. Trønder also has several versions of *mye* in its top 50 features: two with a medial /-t-/ and one without. The only other variant of this word that made it into a top 50 selection is the East Norwegian /myççy/.

Det and *da*

Two other very high-frequency words that are frequently represented by features with high importance scores but usually not discussed as characteristic dialect features are *det* 'it, that, the, there' and *da* 'then' (Table 3.14). The results show some vowel variations in different dialect groups—including notable intra-group variation for West Norwegian, which includes four variants of *det*, each with a different vowel. For both *det* and *da*, the important East Norwegian features include (but are not limited to) variants with an initial /r-/ that replaces the /d-/. While this is generally not described as a typical East Norwegian feature, this resembles the (documented) reduction of /d-/ to /r-/ in second person pronouns in some East Norwegian dialects (Hårstad and Opsahl, 2013, p. 88; Endresen, 1990, p. 97; Lie, 1990, p. 184; see also subsubsection 3.6.3). The lenition of *det* to /e/ in Trønder is also not generally documented in descriptions of the Trønder dialect area (cf. Mæhlum and Røyneland, 2012, pp. 75–85), but this is also similar to a high-ranking pronoun feature where 3.PL *de(m)* is reduced to /æmm/ (see subsubsection 3.6.3).

Retroflexes

In most parts of Norway, a phonological sequence of /r/ followed by a different alveolar consonant undergoes assimilation, resulting in a retroflex consonant. The exception to this is (by and large) West Norway, where no such assimilation happens and where /r/ often is realized as a uvular consonant rather than an alveolar (Mæhlum and Røyneland, 2012, pp. 90, 185). The ScanDiaSyn transcription system does not distinguish between different realizations of /r/ and the only retroflexes it explicitly encodes are /ɽ/ (which is not the result of assimilation, but see subsubsection 3.6.2 for more on this sound) and /ʂ/. Two features encoding the non-assimilation of /rs/ are among the fifty input features with the highest importance scores for West Norwegian: `rs` and `rrs`. The former denotes the sequence of /rs/ in any syllable and the latter more specifically in short, stressed syllables.

	Group	Feature	Imp.	Dist.	Rep.	Context (bokmål/pron.)
det	West	<SOS>det/dær<EOS>	0.19	0.01	0.84	
		<SOS>det/da<EOS>	0.18	0.10	0.96	
		dår	0.13	0.01	0.85	det/dårr(0.4)
		<SOS>det/di<EOS>	0.11	0.01	0.54	
	Trø.	<SOS>det/e<EOS>	0.12	0.03	0.39	
	East	ræ<SEP>	0.13	0.01	0.60	det/ræ(0.9)
		<SEP>re	0.12	0.10	0.46	det/re(0.9)
		<SOS>det/re<EOS>	0.10	0.07	0.93	
		<SOS>det/de<SEP>er/ær<EOS>	0.08	0.05	0.98	
	North	<SOS>det/d<SEP>er/e<EOS>	0.08	0.04	0.44	
		<SOS>det/de<SEP>der/dær<EOS>	0.09	0.01	0.49	
da	West	då<SEP>	0.14	0.16	0.53	da/då(1.0)
	East	<SOS>da/ra<EOS>	0.27	0.02	0.95	
		<SOS>da/a<EOS>	0.09	0.03	0.42	

Table 3.14: Variants of *det* 'it, that, the, there' and *da* 'then' in the top 50 most important features per dialect group. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each character n-gram appears (along with the relative frequency of this word being the origin).

	Group	Feature	Imp.	Rep.	Dist.	Context (bokmål/pron.)
/i/ > /e/	North	vess	0.11	0.01	0.54	hvis/vess(0.6) 'if'
		<SOS>til/ti<EOS>	0.09	0.02	0.22	'to'
		fessk	0.09	0.01	0.60	fisk/fessk(0.2) 'fish'
		vess<SEP>	0.07	0.01	0.70	hvis/vess(1.0) 'if'
		<SEP>tell	0.07	0.01	0.43	til/tell(0.8) 'to'
	Trø.	ekker	0.05	0.01	0.35	sikkert/sekkert(0.8) 'sure(ly)'
/ao∼åo/	West	åo	0.12	0.01	0.72	da/dåo(0.1) 'there'
		ao	0.08	0.02	0.83	au/ao(0.2) 'also; ouch'

Table 3.15: Vowel patterns in the top 50 most important features per dialect group. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each character n-gram appears (along with the relative frequency of this word being the origin).

Vowels

As shown in Table 3.15, several of the high-ranking North Norwegian features encode a sound change that is common to many dialects of that group: the lowering of /i/ to /e/ (Hanssen, 2010, p. 189). This sound change is demonstrated in features for the words *hvis* 'if,' *fisk* 'fish,' and *til* 'to,' although the latter is also represented by a high-importance feature with /i/. This sound change is also typical for Trønder dialects (Hanssen, 2010, p. 157), although only one feature representing this made it into that group's top 50 LIME features.

One diphthong that is characteristic of a few West Norwegian dialects is /ao∼åo/ (Sandøy, 1990, p. 76; Hanssen, 2010, p. 172). Unlike many other dialect traits that are represented by entire words or longer character n-grams in the highest-ranking LIME results, these features only encode the diphthong itself: `ao` and `åo`.

Inflected verb forms

Many of the features with high importance scores represent conjugated forms of common verbs, as shown in Table 3.16. These are often not specifically discussed by dialectologists, but some of them exemplify other dialect traits. For instance, the final /-r/ in unstressed syllables (such as in the present tense forms of many verbs) is dropped in many Norwegian dialects, with the exception of the East Norwegian group Mæhlum and Røyneland (2012, pp. 53, 79, 92, 110). This tendency is also reflected by the entries for *har* 'have.PRES,' *er* 'am, are, is,' and *var* 'was' in Table 3.16.

The variants of *vært* showcase a typical ending of past participle forms in many Trønder and East Norwegian dialects: /-i/ (Dalen, 1990, p. 134; Endresen, 1990, p. 96).

Past participles ending in /-dd/

Four of the North Norwegian features with the highest LIME scores represent past participles ending in /-dd/ instead of the more prevalent /-tt/. (The examples in the context column of Table 3.17 are far from exhaustive. Most of the words in which `dd<SEP>` appears are past participle forms of a broad range of verbs, such as *gått* /gådd/ 'gone,' *fått* /fådd/ 'gotten,' *sett* /sedd/ 'seen,' *hatt* /hadd/ 'had (PST-PCP),' and many others).

In the ScanDiaSyn data, these forms mostly appear in the North Norwegian samples (note the high distinctiveness scores) and most of the North Norwegian utterances include the /-dd/ versions and not the /-tt/ versions (for instance, 87 % of the appearances of *gått* 'gone' are pronounced /gådd/ in the North Norwegian ScanDiaSyn data). Nevertheless, this is not discussed as a characteristic trait of North Norwegian by, e.g., Mæhlum and Røyneland (2012, pp. 109–110).

	Group	Feature	Imp.	Rep.	Dist	Context (bokmål/pron.)
ble(i) 'became'	North	<SEP>bei	0.08	0.01	0.79	blei/bei(0.8)
	East	<SEP>bɽe	0.11	0.01	0.84	ble/bɽe(0.8)
		ɽæi	0.08	0.01	0.87	blei/bɽæi(0.7)
	Trø.	<SOS>ble/varrt<EOS>	0.06	0.03	0.26	
gjør 'do. PRES'	Trø.	<SOS>gjør/jær<EOS>	0.12	0.01	0.35	
	West	<SEP>jer	0.09	0.01	0.45	gjør/jer(0.6)
har 'have. PRES'	Trø.	hi<SEP>	0.23	0.01	0.84	har/hi(1.0)
	West	he<SEP>	0.09	0.05	0.65	har/he(1.0)
er 'am, are, is'	East	<SOS>er/ær<EOS>	0.15	0.10	0.97	
		<SOS>er/æ<EOS>	0.13	0.17	0.61	
		<SEP>er	0.10	0.01	0.71	er/er(1.0)
		<SOS>er/er<EOS>	0.09	0.01	0.83	
		<SOS>så/så<SEP>er/ær<EOS>	0.07	0.01	0.99	
	West	<SOS>er/æ<SEP>det/de<EOS>	0.09	0.01	0.24	
var 'was'	East	<SOS>var/var<EOS>	0.16	0.08	0.79	
		<SEP>var	0.07	0.10	0.45	var/var(0.9)
	Trø.	<SOS>var/va<SEP>nå/nå<EOS>	0.10	0.02	0.41	
vært 'been'	East	vør	0.08	0.02	0.37	vært/være(0.2) vært/væri(0.2)
		øri	0.07	0.01	0.44	vært/væri(0.6)
	Trø.	rri<SEP>	0.05	0.01	0.46	vært/værri(0.4)

Table 3.16: Inflected forms of high-frequency verbs in the top 50 most important features per dialect group. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each character n-gram appears (along with the relative frequency of this word being the origin).

Group	Feature	Imp.	Rep.	Dist.	Context (bokmål/phon.)
North	åddSEP	0.12	0.01	0.82	gått/gådd(0.4) 'gone' fått/fådd(0.4) 'gotten'
	iddSEP	0.08	0.01	0.65	blitt/blidd(0.4) 'become.PST-PCP'
	ddSEP	0.07	0.06	0.35	hadde/hadd(0.2) 'had (PRET)'
	ådd	0.07	0.01	0.72	gått/gådd(0.4) 'gone' fått/fådd(0.4) 'gotten'

Table 3.17: Past participles ending with /-dd/ in the top 50 most important features per dialect group. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word(s) in which each character n-gram appears (along with the relative frequency of this word being the origin).

3.6.4 Discussion

Many of the features that got assigned high importance scores by LIME serve as examples for the linguistic patterns described by dialectologists. However, not all features that are important for the label prediction are easy to understand for humans or fall into easily recognizable feature categories. Additionally, many of the features with high-importance scores showcase linguistic traits that are not often discussed in Norwegian dialectology, such as the different variants of *noe(n)* 'some, somebody, something' or the past participle endings in North Norwegian.

The features that have high importance scores for a dialect group are not always very representative of the entire group (although these exist, such as the West Norwegian /(r)rs/), but sometimes only represent characteristic traits of the dialects spoken in one subregion (e.g. the diphthongs /ao, åo/ in some parts of West Norway). This also results in there sometimes being several seemingly contradictory features that have high importance scores for the same label, such as the West Norwegian first person singular variants /i/, /ei/ and /eg/ that are all among the 50 highest-ranking features for that dialect group. It would be interesting to explore how this might change if the number of input features is restricted further and relatively infrequent features are excluded.

It might also be insightful to examine the importance scores for features that are encoded differently, for instance as sound correspondences between the dialects and a reference doculect.

Furthermore, it would be worthwhile to explore which features have high importance scores and are common in false positives/negatives: are there patterns as to which linguistic features lead the classifier astray?

Chapter 4

Case study: Detecting sexism in French tweets

To analyze explainable machine learning in an applied context as well, I consider the case of automated sexism detection in French tweets. This chapter is structured as follows: first, I introduce the topic of automatic sexism detection and previous approaches to this task (section 4.1). I then describe the dataset I work with (section 4.2). In section 4.3, I present the set-up of the LIME-based experiment (subsection 4.3.1) and its results (subsection 4.3.2). I then describe the specifics of the architecture for the attention-based approach (section 4.4) and present the results in subsection 4.4.2. I discuss the results from both approaches in section 4.5.

4.1 Sexism detection

The increased popularity of systems that can automatically detect abusive speech has led to a recent focus on more specific breakdowns by, e.g., specific groups targeted by hate speech or offensive languages. Poletto et al. (2020) present an overview of corpora for hate speech detection, distinguishing between different types of hate speech, languages and annotation styles.

In the last few years, several datasets and automatic classifiers for sexism detection have been published for English (Jha and Mamidi, 2017; Anzovino et al., 2018; Fersini et al., 2018; Frenda et al., 2019), Spanish (Fersini et al., 2018; Rodríguez-Sánchez et al., 2020), Italian (Fersini et al., 2020) and French (Chiril et al., 2020) data.

Anzovino et al. (2018) experiment with different features and machine learning models for identifying misogynistic tweets written in English and further assigning them to more specific subcategories. The features they try out are n-grams of characters, tokens and part-of-speech tags, the tweet length, the presence of URLs and the number of usernames mentioned, the number of adjectives, and token embeddings. The authors compare SVMs, random forests, naive Bayes classifiers and feed-forward neural networks. They find that both for detecting misogynistic tweets in general and for identifying what kind of misogyny is present in a given tweet, SVMs with

token-level 1-3grams perform best.

Chiril et al. (2020) introduce the French twitter dataset that I also work with, which is described in section 4.2. They compare the performance of different classifiers, including an SVM with token 1-3grams, a bidirectional LSTM with attention and a multilingual BERT model with an additional classification layer. The authors find that the BERT model produces by far the best results. When the input data to the SVM are preprocessed such that URLs are replaced by the title of the website they link to and emoji are replaced with custom descriptions, the SVM clearly outperforms the bi-LSTM with attention. Chiril et al. point out that many prediction errors occur when a tweet includes ironic statements, when additional reasoning or knowledge of the world is required to understand the tweet, or when tweets contain stereotyping statements but do not include swear words or insulting vocabulary.

Frenda et al. (2019) analyzed several English-language corpora of sexist tweets and find that sexist tweets tend to have a lower type-token ratio and contain more swear words and more feminine pronouns than non-sexist tweets. The authors train SVMs to automatically detect sexist tweets and experiment with different ways of encoding the tweets. They find that using character-level 1-7grams or token-level 1-3grams yields good results when only using one kind of feature encoding, but they obtain their best results when combining the two and additionally adding features that encode whether a tweet contains words that are part of different lexicons relating to vulgarity, femininity, sexuality, the human body, and sexist hashtags.

Pamungkas et al. (2020) tested misogyny detection in English, Italian and Spanish tweets. When comparing different classifiers including SVMs, a BERT-based model, and different types of RNNs with and without pretrained embeddings and with and without attention layers, they find that depending on the dataset, SVMs or BERT perform best. The authors also find that the best way of encoding the data for the SVMs depends on the dataset, although encoding the presence of words relating to women and of sexist slurs tends to be useful in general.

4.2 Data

I work with a dataset of French tweets collected by Chiril et al. (2020).[1] The data consist of French tweets that were collected in 2017 and 2018 using a list of keywords. Such keywords include terms referring to gender or traditionally associated with one gender, gendered insults, public figures who are potential victims or perpetrators of sexism and hashtags used when recounting sexist experiences.

The tweets are annotated based on whether or not they contain sexist content. Tweets with sexist content fall into three categories. By far the smallest subgroup consists of *directly* sexist tweets that are addressed to one or more women:

(3) Les filles qui affichent leurs corps partout et qui se disent fière, féministe ou jsp encore quelle connerie; sachez qu'on peut en être fière sans le montrer au monde entier, donc vous plaignez pas de l'image que vous renvoyez
'Girls who display their bodies everywhere and call themselves proud or feminist or I don't know what other nonsense; know that you can be proud of your body without showing it to the entire world, so don't complain about the image you're giving off.'

(4) Assume! Tu fais tout pour faire le buzz et après tu pleures [EMOJI] quand on vient à moitié à poil chez Ardisson , on sait à quoi s attendre , j en ai marre de ces nanas qui n assument pas et font des histoires au nom de leur féminisme à 2 francs !
'Accept it! You do everything to generate buzz and then you cry [EMOJI] If you go to Ardisson['s talk show] while half-naked, you know what to expect, I'm fed up with chicks who don't stand by what they do and make a fuss in the name of their cheapo feminism!'

Tweets with *descriptive* sexist content are not directly addressed to anybody who would be the target of the sexist content, but describe one or more women:

(5) La cuisine pour une femme EST UN DEVOIR NATUREL comme on parlerai de droit naturel. Déjà moi je le dis tout haut: Je n'épouserais pas une femme qui ne sait pas faire la cuisine même si elle est hyper belle ou riche ou encore possède de dizaines de diplômes, juskà ce k'el l'apren
'For a woman, the kitchen IS A NATURAL TASK, like a law of nature. I say this loudly: I'm not gonna marry a woman who cannot cook even if she's super beautiful or rich or has dozens of diplomas until she learns to cook.'

(6) Les femmes elles sont pas crédibles dans la démarche égalité homme/femme parce qu'elles se respectent déjà pas entre elle
'Women aren't credible in their undertaking for equality between men and women because they don't even respect each other.'

Lastly, there are tweets that *report* experiences with sexism. About 80 % of the tweets with sexist content fall into this category; for instance:

[1] Available at `https://github.com/patriChiril/An-Annotated-Corpus-for-Sexism-Detection-in-French-Tweets`.

Dataset	Sexist content			Non-sexist	Total
Chiril et al. (2020)	4,047	direct	45	7,787	11,834
		descriptive	780		
		reporting	3,222		
My subset	3,278			6,388	9,666

Table 4.1: Class distribution in the Twitter corpus by Chiril et al. (2020) and my subset thereof.

(7) Il y a des gens (hommes) qui mettent leur numéro de tél dans leur bio pour des raisons pro Moi j'ai dû enlever mon numéro de téléphone d'un de mes CV en ligne parce qu'un élève m'a dragué par sms, et un autre mec s'en est servi alors q j'avais refusé de lui donner mon numéro
'There are people (men) who put their phone numbers into their bio for job reasons. I had to remove my phone number from one of my online CVs because a student tried to pick me up via SMS, and another guy used it when I had refused to give him my number.'

(8) La bonne réponse est la réponse D - 1 femme sur 5 sera victime d'un viol ou d'une tentative de viol au cours de sa vie (Source : @USAID). #Ilesttemps de mettre fin aux #VFS #MoiAussi [URL]
'The correct answer is number D—one out of five women becomes a victim of rape or attempted rape in the course of her life (source: @USAID). #TimesUp for putting an end to #GenderBasedViolence #MeToo [URL]

(9) «Si tu veux pas m'offrir ton corps, je peux le louer ?» Bordeaux — place de la Victoire #payetashnek
'"If you don't want to offer your body to me, can I rent it?" Bordeaux—Place de la Victoire #payetashnek[2]'

In the version of the corpus that is available, no fine-grained distinction is made between these three subtypes of the class of tweets with *sexist content* (as of writing this thesis). The remaining tweets are labelled as *non-sexist*.

The publicly available corpus does not contain the full tweets but the list of tweet IDs which can be used to retrieve the posts from Twitter. I retrieved the tweets on December 5, 2020. A portion of the tweets had already been deleted, leaving about 9.700 tweets, about a third of which contain sexist content. Table 4.1 shows the class distribution of the original dataset and the tweets I was able to retrieve.

It should be noted that many of the tweets are written in colloquial French and include texting abbreviations and/or spelling mistakes:

[2]A hashtag used for reporting sexual harassment in public spaces.

(10) **Tweet** Mtn y'a du sexisme envers les
Standard French Maintenant il y a du sexisme envers les

hommes laissez moi rire <URL>
hommes, laissez-moi rire <URL>

'Now there's sexism against men, I got to laugh <URL>'

(11) **Tweet** Mskn tjr je m'excuse, tjr je
Standard French Meskine, toujours je m'excuse, toujours je

pardonne tlm, tjr j'fais le premier pas jsuis
pardonne tout le monde, toujours je fais le premier pas, je suis

trop conne <URL>
trop conne <URL>

'Poor me, I always say sorry, I always forgive everybody, I always take the first stop, I'm too stupid'

(12) **Tweet** Jentends un mec qui dit cest les risques
Standard French J'entends un mec qui dit c'est les risques

du metier que natalie portman ait recu des lettres la
du métier que Natalie Portman ait reçu des lettres la

menaçant de viol A 13 ANS et que lui il a reçu des
menaçant de viol A 13 ANS et que lui, il a reçu des

lettres dinsultes jui MOR quel rappor
lettres d'insultes, je suis MORT(E), quel rapport

'I can hear a guy who is saying that Natalie Portman receiving rape threats as a THIRTEEN YEAR OLD is a job hazard, and he's also received letters with insults, I'm DEAD, what a comparison'

4.2.1 General preprocessing

I preprocess the Twitter data by replacing usernames, numbers and hashtags with `<USER>`, `<NUMBER>` and `<HASHTAG>`, respectively. I replace URLs with the title of the linked website when available, and remove them otherwise. Chiril et al. (2020) found that replacing URLs led to better classification results. I also normalize the punctuation by mapping different kinds of apostrophes and quotation marks to standard versions thereof.

4.3 LIME

4.3.1 Preprocessing and method

Instead of encoding the tweets as character- and word-level n-grams as with the dialect data, I use the sub-word tokenization produced by the tokenizer of the (cased, large) FlauBERT model (Le et al., 2020), a pre-trained BERT model for text written in French. This method encodes frequent words as word unigrams and less frequent words as subword units, based on byte pair encoding (BPE). Unlike the encoding based on n-grams of different lengths, there is no overlap between tokens. Tokenizing the input like this led to an improvement in the model accuracy and F_1-score in preliminary experiments, and it produces features that are more easily interpretable for humans. The example below gives an impression of what the features encoding a tweet after preprocessing look like. Each (non-escaped) token ends with a hyphen or with </w>:

(13)

Tweet	#griveaux	#hulot	ou	le	retour
Encoded	<HASHTAG>	<HASHTAG>	ou</w>	le</w>	retour</w>

des	“pater	familias”	autant
des</w>	"</w> pa- ter</w>	famili- as</w> "</w>	autant</w>

dire	la	négation	radicale	du
dire</w>	la</w>	négation</w>	radicale</w>	du</w>

féminisme	par	ces	spécialistes	de
féminisme</w>	par</w>	ces</w>	spécialistes</w>	de</w>

l’égalité	femme-homme,	qu’en
l’</w> égalité</w>	femme</w> -</w> homme</w> ,</w>	qu’</w> en</w>

pense	Marlène	Schiappa	?
pense</w>	Marl- ène</w>	Schi- appa</w>	?</w>

‘#Griveaux, #Hulot,[3] or the return of the “pater familias,” which is to say the radical negation of feminism by these gender equality specialists—what does Marlène Schiappa[4] think about this?’

In a preliminary experiment, I tried out encoding the data as word- and character-level n-grams (similarly to my approach to the dialect classification task and to the approaches listed in the first section of this chapter). However, the BPE-based tokenization yielded slightly better classification results (all other settings being equal) as well as features that are much easier to understand for humans.

The machine learning model I use is an SVM, since this kind of model has proven to perform well in many of the experiments mentioned in section 4.1. As in the dialect experiment, I create ten different train-test splits of the data, train an SVM on each of them, and average the accuracy and F_1 metrics as well as the LIME scores across these ten set-ups. I also use TF-IDF weighting for numerically representing the input

[3]Benjamin Griveaux and Nicolas Hulot are French politicians.

[4]Marlène Schiappa was the French Secretary of State for Gender Equality from 2017 to 2020.

tokens, including the 5000 most common tokens. Because the label distribution is clearly imbalanced, I use class weights (giving twice the weight to tweets with sexist content while training).

4.3.2 Results

The SVMs have an average accuracy of 77.0 % and an mean (macro-averaged) F_1-score of 74.8 %.

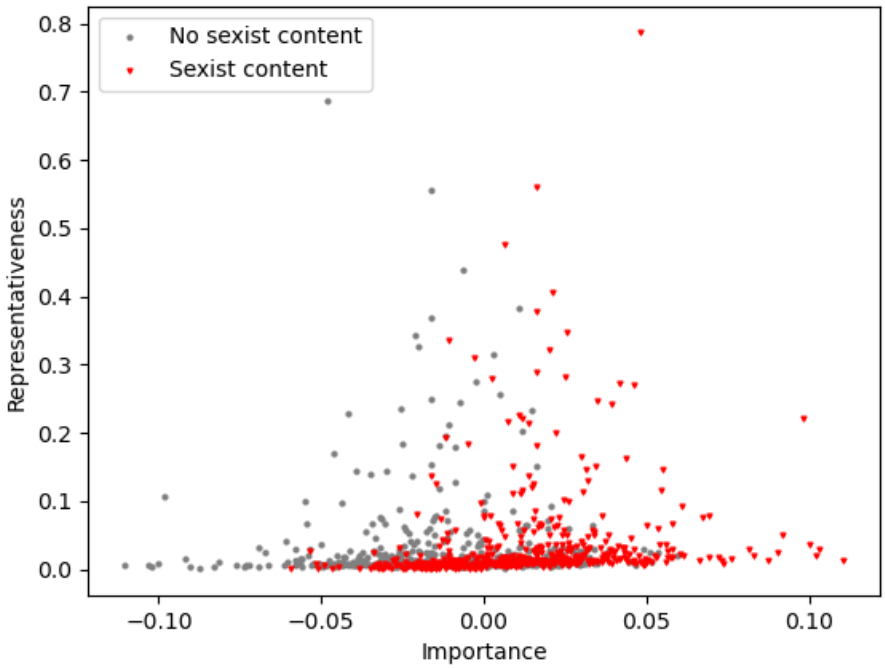


Figure 4.1: Representativeness values by LIME importance scores.

As with the dialect experiment, the importance scores and representativeness scores are (almost) independent of one another (Figure 4.1). The correlation efficient between the two scores is 0.03 for importance values both for tweets with and without sexist content. Most features that appear especially often in either of the tweet types have importance scores that are close to zero.

By contrast, as Figure 4.2 shows, the distinctiveness scores and the LIME importance values show a clear correlation: the higher the importance score of a feature is for a given label, the more specific it is to samples with that (gold standard) label. For both labels, the correlation coefficient for importance and distinctiveness is 0.8. Importance scores for the group of tweets with sexist content are generally slightly higher than those for features in non-sexist tweets (see also Figure 4.2).

Tables 4.2 and 4.3 show the local importance scores for the tweet from Example 13. Since this is a binary classification task, each feature has the same absolute importance scores for both labels—only the sign is inverted.

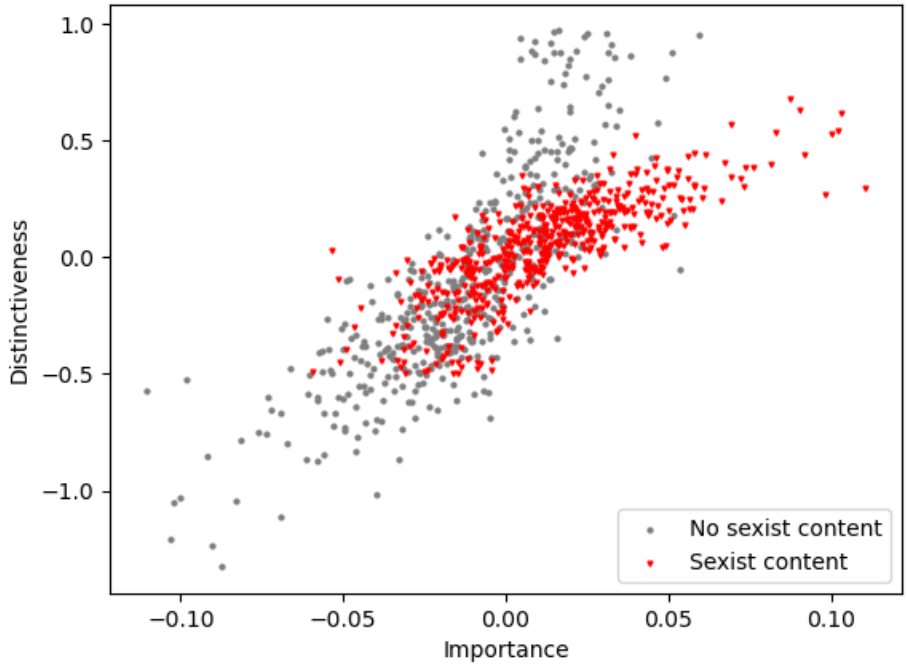


Figure 4.2: Distinctiveness values by LIME importance scores.

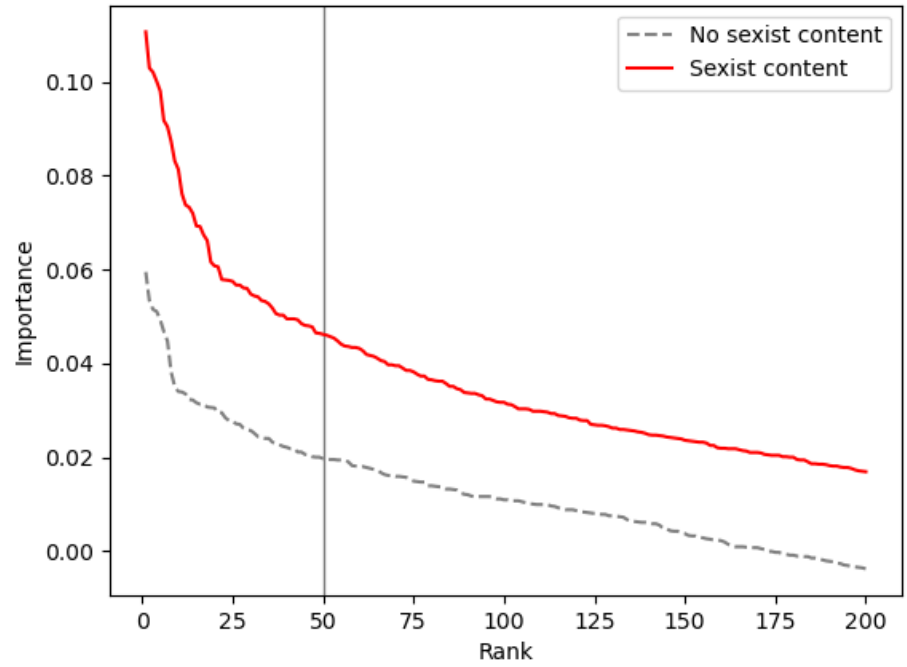


Figure 4.3: Importance scores for the 200 highest-ranking tweets per label.

		Sexist content		No sexist content	
<HASHTAG>		0.07	<HASHTAG>	-0.07	<HASHTAG>
<HASHTAG>		0.07	<HASHTAG>	-0.07	<HASHTAG>
ou	‘or’				
le	‘the’				
retour	‘return’				
des	‘of the’				
"		0.07	"</w>	-0.07	"</w>
pater		0.05	pa-	-0.05	pa-
familias					
"		0.07	"</w>	-0.07	"</w>
autant	‘as much as’	-0.05	autant</w>	0.05	autant</w>
dire	‘saying’				
la	‘the’				
négation	‘negation’				
radicale	‘radical’				
du	‘of the’				
féminisme	‘feminism’	0.11	féminisme</w>	-0.11	féminisme</w>
par	‘by’				
ces	‘these’	0.05	ces</w>	-0.05	ces</w>
spécialistes	‘specialists’				
de	‘of’				
l’	‘the’				
égalité	‘equality’	0.10	égalité</w>	-0.10	égalité</w>

Table 4.2: Local importance scores for a sample tweet. Features with importance scores between -0.05 and +0.05 are omitted to preserve space. (Continued in the following table.)

		Sexist content		No sexist content	
femme	‘woman’	0.11	femme</w>	-0.11	femme</w>
-					
homme	‘man’	0.08	homme</w>	-0.08	homme</w>
,					
qu’	‘what’				
en	‘of it’				
pense	‘thinks’				
Marlène					
Schiappa					
?					

Table 4.3: (Continuation of the previous table.) Local importance scores for a sample tweet. Features with importance scores between -0.05 and +0.05 are omitted to preserve space.

In the following, I examine patterns into which the top 50 most important features per label fit. This cut-off point was also chosen to include most of the features with comparatively high importance scores while still retaining a large enough group in order to find patterns within this sub-selection. Figure 4.3 shows the importance scores for the highest-ranking features in both label classes. The 200 highest-ranking features per class can be found at `https://github.com/verenablaschke/ma-thesis/tree/main/models/tweets`.

Words relating to gender

Many of the tokens that appear in tweets with sexist content and that have high importance scores relate to gender in some way (Table 4.4). Most of these are words or subtokens of words that describe women (*filles* ‘girls,’ *femme(s)* ‘woman/women,’ *meuf(s)* ‘woman/women (colloq.)’ and *elles* ‘they.FEM’),[5] although two words for men also make it into the top 50 (*homme* ‘man,’ *mec* ‘guy’). Notably, the class of tweets without sexist content also contains one high-importance token referring to women, which is the singular form of one of the aforementioned features: *fille* ‘girl.’ However, this word actually appears marginally less often in tweets with this label as one would a randomly distributed feature expect to appear.

[5]The latter does not always refer to groups of women, it substitutes any plural noun phrase that is grammatically feminine.

Label	Feature	Imp.	Rep.	Dist.	Context
Sexist content	filles</w>	0.10	0.02	0.54	filles(1.0) 'girls'
	femme</w>	0.10	0.22	0.27	femme(1.0) 'woman'
	mec</w>	0.09	0.02	0.63	mec(1.0) 'guy'
	meu-	0.09	0.01	0.68	meuf(0.7) 'woman'
	homme</w>	0.07	0.08	0.41	homme(1.0) 'man'
	femmes</w>	0.05	0.12	0.20	femmes(1.0) 'women'
	elles</w>	0.05	0.02	0.34	elles(0.9) 'they.FEM'
	Femme</w>	0.05	0.01	0.15	Femme(1.0) 'woman'
No sexist content	fille</w>	0.05	0.03	-0.05	fille(1.0) 'girl'

Table 4.4: Features with the top 50 highest LIME scores per label that are related to gender. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each token appears (along with the relative frequency of this word being the origin).

Words relating to feminism and sexism

Several of the features with high importance scores are connected to discussions of feminism or sexism. Here, tokens relating to feminism or equality are indicators of tweets with sexist content, while the features mentioning sexism are split: sexiste</w> 'sexist' is an indicator for sexist content whereas sexisme</w> 'sexism' has a high importance score for tweets with *no* sexist content (Table 4.5).

Gendered insults

Two of the high-ranking features of tweets with sexist contents are (subtokens of) insults directed at women: salope</w> 'slut, bitch' and asse</w>, which almost always appears as the suffix of *connasse* 'bitch' in this dataset. However, one of the top 50 indicators of a non-sexist tweet is also such a term: *conne* 'bitch; stupid.FEM' Both of the features with high importance scores for the sexist class have high distinctiveness scores, whereas *conne* only occurs in the tweets without sexist content roughly as often as one would expect by chance (see Table 4.6).

The presence of insults that either relate to intellectual deficits or that pertain to sexuality fits with the observations that Dupré and Gramaccia (2020) made in a qualitative analysis of sexist tweets.

Label	Feature	Imp.	Rep.	Dist.	Context
Sexist content	féminisme</w>	0.11	0.01	0.30	féminisme(1.0) 'feminism'
	égalité</w>	0.08	0.03	0.40	égalité(1.0) 'equality'
	sexiste</w>	0.06	0.02	0.25	sexiste(1.0) 'sexist'
	féministe</w>	0.06	0.03	0.21	féministe(1.0) 'feminist'
No sexist content	sexisme</w>	0.03	0.06	0.35	sexisme(1.0) 'sexism'

Table 4.5: Features with the top 50 highest LIME scores per label that are directly related to feminism or sexism. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each token appears (along with the relative frequency of this word being the origin).

Label	Feature	Imp.	Rep.	Dist.	Context
Sexist content	salope</w>	0.08	0.02	0.53	salope(1.0) 'slut, bitch'
	asse</w>	0.08	0.02	0.38	connasse(0.8) 'bitch'
No sexist content	conne</w>	0.03	0.01	0.02	conne(1.0) 'stupid.FEM'

Table 4.6: Features with the top 50 highest LIME scores per label that are (subtokens) of insults directed at women. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each token appears (along with the relative frequency of this word being the origin).

Label	Feature	Imp.	Rep.	Dist.	Context
No sexist content	Ségolène</w>	0.05	0.03	0.88	Ségolène(1.0)
	Christiane</w>	0.05	0.03	0.77	Christiane(1.0)
	Royal</w>	0.04	0.03	0.87	Royal(1.0)
	Angela</w>	0.03	0.06	0.86	Angela(1.0)
	Theresa</w>	0.03	0.03	0.91	Theresa(1.0)
	Christine</w>	0.03	0.02	0.77	Christine(1.0)
	Taubira</w>	0.03	0.03	0.74	Taubira(1.0)
	Lagarde</w>	0.02	0.01	0.78	Lagarde(1.0)
	May</w>	0.02	0.03	0.89	May(1.0)

Table 4.7: Features with the top 50 highest LIME scores per label that are names of female politicians. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each token appears (along with the relative frequency of this word being the origin).

Female politicians

Many of the features with high importance scores for the tweets without sexist content are the first and last names of several influential female politicians (Ségolène Royal, Christiane Taubira, Christine Lagarde, Angela Merkel and Theresa May). While none of these names appear in an especially high proportion of the non-sexist tweets, all them them appear almost exclusively in this class (Table 4.7).

Pronouns

As Table 4.8 shows, ten of the 50 features with the highest importance scores for the tweets with sexist contents are first and second person singular (informal) pronouns, and these features do indeed also have high distinctiveness scores for that class. No personal pronouns are included in the top 50 features for the other label. The fact that so many first and second person pronouns are indicators for sexist content fits the definition of two of the subcategories within that class in the corpus: tweets with sexist content that is *directed* at someone often include second person pronouns by their nature, and tweets *reporting* encounters with sexism often contain first person pronouns (or second person pronouns, if they include direct quotations).

Punctuation

Four of the features that LIME deems indicative of sexist content consist of punctuation marks (-</w>, . . .</w>, =</w> and ?-), as does one of the high-ranking features

Label	Feature	Imp.	Rep.	Dist.	Context
	t’</w>	0.10	0.04	0.53	t’(1.0) ‘your.SG’
	te</w>	0.09	0.05	0.44	te(0.8) ‘you.SG.ACC’
	ta</w>	0.06	0.02	0.26	ta(0.9) ‘your.SG’
	tes</w>	0.06	0.02	0.45	tes(0.6) ‘your.SG’
Sexist	ton</w>	0.06	0.03	0.31	ton(1.0) ‘your.SG’
content	me</w>	0.06	0.07	0.31	me(0.9) ‘me’
	moi</w>	0.06	0.04	0.34	moi(1.0) ‘I, me’
	ma</w>	0.05	0.04	0.25	ma(1.0) ‘my’
	toi</w>	0.05	0.02	0.37	toi(1.0) ‘you’
	mes</w>	0.05	0.01	0.16	mes(0.9) ‘my’

Table 4.8: Features with the top 50 highest LIME scores per label that are personal pronouns. The middle columns contain importance, representativeness and distinctiveness scores. The context column lists the most frequent word in which each token appears (along with the relative frequency of this word being the origin).

of the other class (,-).[6]

[6]This should not be confused with the more common comma feature ,</w>. The above-mentioned feature is much rarer and appears in character sequences such as ," where it is followed by other letters (usually other punctuation marks).

4.4 Attention

4.4.1 Preprocessing and method

I lowercase the tweets and embed them using pretrained word2vec (Mikolov et al., 2013) embeddings for French, as provided by Fares et al. (2017).[7] These embeddings work on a word (and punctuation) level rather than a sub-token level. I truncate all tweets that are longer than sixty tokens (that is, words or (clusters of) punctuation marks) and pad all shorter tweets with dummy tokens (<FILLER>).

I use a feed-forward neural network with an attention layer, as illustrated in subsection 2.2.1. In preliminary experiments, the choice between an FFNN and a recurrent neural network did not lead to significant differences in the classification accuracy. I therefore use the FFNN since the hidden representations produced by a recurrent model might be less directly reflective of the individual input tokens. I use a hidden layer size of 128 for the FFNN, a dropout rate of 0.4 between the FFNN and the attention layer, and train the model with a batch size of 64 and a learning rate of 0.01 with an Adam optimizer. To build and train the model, I use the Python library Keras[8] 2.4.3 with a Tensorflow[9] 2.4.1 backend.

As with the other experiments, all metrics and scores are averaged across ten initializations and train-test splits.

4.4.2 Results

The neural classifiers have an average accuracy of 73.9 % and F_1 score of 72.5 % across the ten initializations.[10]

I extract attention weights for tokens in the test sets and discard those that appear less than twenty times. I then calculate the global attention score for each token by averaging the attention weights that the token is associated with in the different tweets and initializations. The highest global attention score is 0.27.

I also examine the distribution of the attention weights per utterance. On average, the entropy of the attention weight distribution for a tweet is 2.58, with a standard deviation of 1.24. For comparison, the maximum possible entropy for a probability distribution with 60 possible outcomes is 4.09. Some tweets have attention weight vectors that are very nearly one-hot encoded (with an entropy of 0.0001), i.e. where

[7] They can be downloaded via `http://vectors.nlpl.eu/repository/20/43.zip`.

[8] `https://keras.io`

[9] `https://www.tensorflow.org/`

[10] These scores cannot be directly compared to the classifier performances that Chiril et al. (2020) report for the same dataset, since their test set has a different label distribution than mine. The test sets I work with all have a label ratio of approximately 1:2 (sexist content vs. non-sexist content), whereas Chiril et al. use a more balanced ratio of circa 3:5. That said, their neural, non-BERT models achieve accuracy scores of up to 69.5 % and F_1 scores of up to 64.0 %. Their best model is a mulitilingual BERT model (Devlin et al., 2019) with a classification layer that has an accuracy of 79.0 % and an F_1 score of 76.2 %.

the attention lies very clearly on a single token, whereas some others have uniform attention distributions (with an entropy of 4.09), but most lie somewhere in the middle.

The <FILLER> tokens have a mean attention weight of 0.01, i.e. an attention score that is only marginally higher than the weight of 0 one might expect.

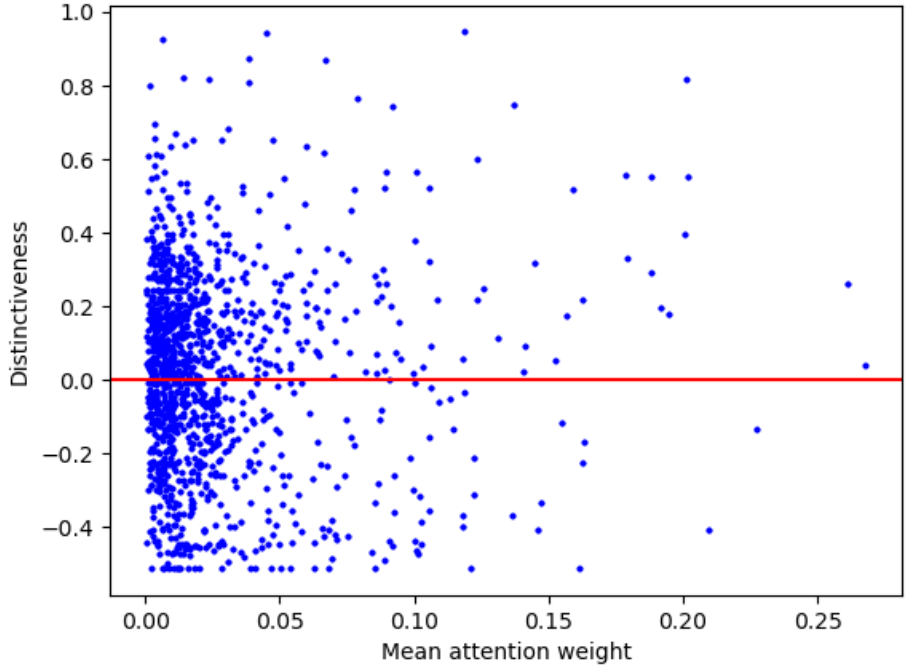


Figure 4.4: Distinctiveness values by global attention weight.

Figure 4.4 shows the global attention weights as well as the corresponding distinctiveness scores. The latter are calculated with regard to the class of tweets with sexist content. Positive distinctiveness scores indicate that a feature appears especially often in sexist tweets (the upper bound is 1.0) and negative scores indicate that a feature occurs especially often in non-sexist tweets (with a lower bound of -0.51). Unlike in the LIME experiments, there is no significant correlation between the attention(/importance) score a feature has and how distinctive it is: many features that are very characteristic of one class of tweets have low attention weights, and some of the features with high global attention scores have distinctiveness values close to zero. For instance, the feature with the highest attention score, `pourrait` 'could,' has a distinctiveness score of only 0.04.

In the rest of this section, I consider the 100 tokens with the highest global attention weights (this represents a range of attention scores from 0.08 to 0.27) and compare recurrent types of tokens present in that group to those present in the LIME results (subsection 4.3.2). The full list of attention weights is available at `https://github.com/verenablaschke/ma-thesis/tree/main/models/tweets-attn`.

The attention weights are not label-specific—a high attention weight only means that the FFNN-encoded representation of a token receives a greater weight when making

the final classification decision. I therefore consider all (high-attention) features with positive distinctiveness scores to be indicators of sexist content, and features with negative distinctiveness scores to be important predictors for tweets with no sexist content.

Words relating to gender or sex

Group	Feature	Imp.	Rep.	Dist.	LIME?
	`filles` 'girls'	0.20	0.02	0.55	`filles</w>`
	`sexes` 'sexes'	0.20	0.01	0.39	
	`femmes` 'women'	0.19	0.12	0.20	`femmes</w>`
	`sexe` 'sex'	0.16	0.01	0.18	
Sexist	`garçons` 'boys'	0.12	0.01	0.60	
content	`dame` 'lady'	0.10	0.00	0.02	
	`féminin` 'feminine, female'	0.09	0.01	0.07	
	`mecs` 'guys'	0.09	0.01	0.74	
	`fille` 'girl'	0.09	0.03	0.03	`fille</w>`
	`femme` 'woman'	0.09	0.23	0.26	`femme</w>`

Table 4.9: Features with the top 100 highest attention scores relating to gender or sex. The middle columns contain importance, representativeness and distinctiveness[11] scores. The right-most column lists the corresponding features presented in subsection 4.3.2, if applicable.

Several of the words with high attention scores relate to gender or sex. They are shown in Table 4.9. Notably, all of these tend to occur especially often in tweets with sexist content (with the exception of *dame* 'lady,' which has a distinctiveness score that is barely above 0). There is some overlap between this group and the gender-related words among the tokens with high LIME importance scores, but many of these terms appear only in the results of one experiment but not the other. These differences *cannot* be explained because of the vocabulary of the word2vec embeddings, since these also contain more colloquial terms like *meuf* 'woman' (which has a high LIME score).

Group	Feature	Imp.	Rep.	Dist.	LIME?
Sexist content	féminisme 'feminism'	0.26	0.01	0.26	
	féministe 'feminist'	0.19	0.03	0.18	féministe</w>
	égalité 'equality'	0.18	0.01	0.33	égalité</w>
	sexistes 'sexist.PL'	0.16	0.01	0.22	
	parité 'parity'	0.14	0.00	0.32	
	féministes 'feminist.PL'	0.14	0.01	0.09	
	sexiste 'sexist'	0.13	0.02	0.25	sexiste</w>
No sexist content	sexisme 'sexism'	0.16	0.04	-0.17	sexisme</w>

Table 4.10: Features with the top 100 highest attention scores relating to feminism or sexism. The middle columns contain importance, representativeness and distinctiveness scores. The right-most column lists the corresponding features presented in subsection 4.3.2, if applicable.

Words relating to feminism and sexism

As in the LIME results, many high-ranking words are directly related to feminism or sexism (Table 4.10). The corresponding features with high attention weights contain all of the feminism/sexism-related tokens with high LIME importance scores, as well as one additional term (*parité* 'parity').

Gendered insults

Three of the tokens with high attention weights are insults directed at women: *connasse* 'bitch' and *salope* 'slut, bitch' (both of which have (subtokens that have) high LIME importance scores for the class of sexist tweets) and *pute* 'whore' (which—unsurprisingly—almost exclusively appears in tweets with sexist content, but is not among the 50 features with the highest LIME importance scores for that class).

Female politicians

Similarly to the LIME results, several tokens with high attention scores refer to female politicians. There is only partial overlap between the two experiments' results however, and the attention-based results also include (female) job titles in addition to names of politicians (Table 4.9). With the exception of the last name of Marlène

[11] Some of the distinctiveness scores deviate slightly from the corresponding ones in the LIME results due to the different tokenization approaches.

Group	Feature	Imp.	Rep.	Dist.	LIME?
No sexist content	députées 'government representatives.FEM'	0.21	0.00	-0.41	
	taubira	0.10	0.01	-0.39	Taubira</w>
	chancelière 'chancellor.FEM'	0.10	0.00	-0.47	
	theresa	0.10	0.00	-0.46	Theresa</w>
	schiappa	0.10	0.00	-0.01	

Table 4.11: Features with the top 100 highest attention scores relating to female politicians. The middle columns contain importance, representativeness and distinctiveness scores. The right-most column lists the corresponding features presented in subsection 4.3.2, if applicable.

Schiappa (who used to be the French Secretary of State for Gender Equality), all of these tokens mostly appear in tweets without sexist content.

Pronouns and punctuation

Unlike the results of the LIME experiment, none of the high-ranking tokens are personal pronouns or contain punctuation marks.

Body parts

Group	Feature	Imp.	Rep.	Dist.	LIME?
Sexist content	seins 'breasts'	0.20	0.01	0.82	
	bite 'dick'	0.14	0.01	0.75	
	fesses 'buttocks'	0.08	0.01	0.77	

Table 4.12: Features with the top 100 highest attention scores describing body parts. The middle columns contain importance, representativeness and distinctiveness scores. The right-most column lists the corresponding features presented in subsection 4.3.2, if applicable.

Three of the tokens with high global attention weights refer to body parts, as shown in Table 4.12. All of these words mostly appear in tweets with sexist content, and none of them are among the tokens with the highest LIME importance scores.

4.5 Discussion

Recurring types of features

Both explanation approaches yield recurring types of features among the tokens with the highest importance or attention scores (although not every high-ranking feature fits into such a group). Despite the differences in how the importance and attention scores are obtained and despite the fact that the LIME importance scores show a high correlation with distinctiveness scores, which the global attention weights do not, the recurring types of features in both experiments' results are very similar.

Some of these features (gendered insults) appear like very plausible indicators of sexist content in a text, whereas others (some types of punctuation marks) are very opaque (even if they mostly appear in only one class of tweets in the dataset). Many of the features with high importance scores or attention weights make sense in the context of the dataset. It is not surprising if a tweet with sexist content contains words relating to women and/or men, if it explicitly mentions sexism or feminism or (in the case of tweets in the *direct sexism* or *reporting sexism* subgroups) if it contains personal pronouns. However, the presence of any of these features can hardly indicate that any given tweet from outside the dataset has sexist content. Likewise, it seems unlikely that female politicians are never the target of or mentioned in sexist tweets. Even so these kinds of features have high distinctiveness scores and therefore reflect actual patterns in the training and test data and therefore, this information can be valuable when inspecting what a model has learned, in order to know whether it should be used in real contexts, whether a training dataset ought to be enlarged or what kinds of unconvincing features with high importance values should be masked when encoding the data.

Tokenization

While the FlauBERT-based tokenization yields mostly easy-to-understand tokens and sometimes produces subtokens that are useful representations for several similar words (e.g. `meu-` for *meuf* 'woman (colloq.)' and *meufs* 'women (colloq.)'), the tokenization might be improved if it split tokens slightly more often and treated subtokens at the beginning/middle and at the end of a word identically. That is, representing for instance *filles* as `fille` and `s` such that the first subtoken also represents the singular form *fille*, might result in interesting generalizations over how related forms of a lemma are used. To then capture possibly relevant effects of inflection or derivation, token bigrams could additionally be used.

It would be interesting to repeat the attention experiment with a similar tokenization, to make the results of both approaches more immediately comparable and inspect the correlation between LIME importance scores and global attention weights.

Attention

Unlike the LIME scores, the global attention weights do not show a strong correlation with distinctiveness, which means that they are a lot less reflective of what features are characteristic of each of the tweet classes. This fits with the discussions presented in subsection 2.2.3 that argue that attention weights do not necessarily provide reliable explanations. Furthermore, this makes the results of the attention-based approach less trustworthy than LIME when considering how well this output might reflect what the model has learned. After all, the neural model's classifications are not that much worse than the SVM's, despite the apparent focus on not very distinctive features and lack of focus on many tokens with high distinctiveness scores.

In this experiment, I use non-contextual token embeddings and a feed-forward neural network. This yields classification results that are slightly worse than the results of the SVM architecture, although they are still comparable. I used a FFNN as neural encoder since in such an architecture, the attention weights appear to correspond more closely to the input tokens than in other encoders (see the discussion in subsection 2.2.3). Even so, contextual embeddings and recurrent neural networks are generally used more commonly and closer to the state of the art, which would make it interesting to re-run this experiment with such a set-up and compare the results. In that case it would be especially interesting to investigate whether switching to a recurrent encoder significantly affects the features' global attention weights.

In future experiments, it would be also be interesting to compare the global attention weights to the global importance scores produced using the method by Ribeiro et al. (2016) that I mention in subsection 2.1.2. That method for generating global scores from the local attention weights works independently of the instance label, as do the global attention weights.

I use an SVM for the LIME-based approach, since that architecture has proven to be suitable for detecting sexist tweets. However, it would be interesting to also extract LIME importance scores from the attention architecture and directly compare the attention and importance scores that stem from one and the same model.

Lastly, it might also be insightful to examine the cases in which the attention weights were (nearly) one-hot encoded and the tweets that received (near-)uniform attention weight distributions.

Chapter 5

Conclusion

Both in a traditional linguistic context such as dialectology and in a more recent applied context such as detecting sexist content in tweets, applying explainable machine learning techniques can be insightful. In both tasks, many of the features with high importance scores fall into recurring groups that can be analyzed. In the case of dialect classification, some of these groups fit in with common dialectological observations while others present patterns in the data that are not typically discussed. In the context of tweet classification, these groups of features can be used to determine whether a model is trustworthy enough to be used in real applications. While the attention weights for input features produce somewhat similar high-ranking results as LIME does, they fail at putting particular focus on highly distinctive features and thus produce less trustworthy insights into the model's classification process.

There is ample opportunity for continuing this work, for instance by exploring which input features with high importance scores tend to appear in incorrectly classified utterances or tweets, or by applying other kinds of explainable machine learning and comparing the results.

Bibliography

Anzovino, M., E. Fersini, and P. Rosso (2018). Automatic identification and classification of misogynistic language on Twitter. In *Proceedings of the 23rd International Conference on Applications of Natural Language to Information Systems (NLDB 2018)*, pp. 57–64. Springer.

Baehrens, D., T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller (2010). How to explain individual classification decisions. *Journal of Machine Learning Research 11*(61), 1803–1831.

Bahdanau, D., K. Cho, and Y. Bengio (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio and Y. LeCun (Eds.), *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*.

Barðdal, J., N. Jörgensen, G. Larsen, and B. Martinussen (1997). *Nordiska: Våra språk förr och nu*. Studentlitteratur.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion 58*, 82–115.

Beijering, K., C. Gooskens, and W. Heeringa (2008). Predicting intelligibility and perceived linguistic distance by means of the levenshtein algorithm. *Linguistics in the Netherlands 25*(1), 13–24.

Bestgen, Y. (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)*, Valencia, Spain, pp. 115–123. Association for Computational Linguistics.

Chakrabarty, T., K. Gupta, and S. Muresan (2019). Pay "attention" to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy, pp. 70–79. Association for Computational Linguistics.

Chakravarthi, B. R., M. Găman, R. T. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, R. Priyadharshini, C. Purschke, E. Rajagopal, Y. Scherrer, and M. Zampieri (2021). Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2021)*, Kiyv, Ukraine, pp. 1–11. Association for Computational Linguistics.

Chiril, P., V. Moriceau, F. Benamara, A. Mari, G. Origgi, and M. Coulomb-Gully (2020). An annotated corpus for sexism detection in French tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pp. 1397–1403.

Çöltekin, Ç. (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*, Barcelona, Spain (Online), pp. 186–192. International Committee on Computational Linguistics (ICCL).

Çöltekin, Ç., T. Rama, and V. 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)*, Santa Fe, New Mexico, USA, pp. 55–65. Association for Computational Linguistics.

Dalen, A. (1990). Dei trønderske dialektane [The *Trønder* dialects]. See Jahr (1990a), pp. 119–139.

Devlin, J., M.-W. Chang, K. Lee, and K. 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)*, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics.

Dupré, D. and G. Gramaccia (2020). Violences de genre et discours post-féministes sur Twitter: Le cas de l'affaire Orelsan [Gendered violence and post-feminist discourses on Twitter: The case of Orelsan]. *Les Enjeux de l'Information et de la Communication 21*(1), 91–112.

Endresen, R. T. (1990). Vikværsk–målet i Østfold, Vestfold, Grenland og Nedre Buskerud [Vikværsk–the language of Østfold, Vestfold, Grenland and Lower Buskerud]. See Jahr (1990a), pp. 89–99.

Fares, M., A. Kutuzov, S. Oepen, and E. Velldal (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa)*, Gothenburg, Sweden, pp. 271–276. Linkøping University Electronic Press.

Fersini, E., D. Nozza, and P. Rosso (2020). AMI @ EVALITA2020: Automatic misogyny identification. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*.

Fersini, E., P. Rosso, and M. Anzovino (2018). Overview of the task on automatic misogyny identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pp. 214–228.

Frenda, S., B. Ghanem, M. Montes-y-Gómez, and P. Rosso (2019). Online hate speech against women: Automatic identification of misogyny and sexism on Twitter. *Journal of Intelligent & Fuzzy Systems 36*(5), 4743–4752.

Garreau, D. and U. von Luxburg (2020). Explaining the explainer: A first theoretical analysis of LIME. In S. Chiappa and R. Calandra (Eds.), *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, Volume 108 of *Proceedings of Machine Learning Research*, pp. 1287–1296. PMLR.

Gooskens, C. (2005). How well can Norwegians identify their dialects? *Nordic Journal of Linguistics 28*(1), 37–60.

Gooskens, C. and W. Heeringa (2006). The relative contribution of pronunciational, lexical, and prosodic differences to the perceived distances between Norwegian dialects. *Literary and Linguistic Computing 21*(4), 477–492.

Găman, M., D. Hovy, R. T. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, C. Purschke, Y. Scherrer, and M. 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 2020)*, Barcelona, Spain (Online), pp. 1–14. International Committee on Computational Linguistics (ICCL).

Hanssen, E. (2010). *Dialekter i Norge [Dialects in Norway]*. Bergen: Fagbokforlaget.

Hårstad, S. and T. Opsahl (2013). *Språk i byen: Utviklingslinjer i urbane språkmiljøer i Norge [Language in the city: Trends in urban linguistic environments in Norway]*. Oslo: Fagbokforlaget.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Ph. D. thesis, University of Groningen.

Heeringa, W. and C. Gooskens (2003). Norwegian dialects examined perceptually and acoustically. *Computers and the Humanities 37*(3), 293–315.

Heeringa, W., K. Johnson, and C. Gooskens (2009). Measuring norwegian dialect distances using acoustic features. *Speech Communication 51*(2), 167–183.

Jahr, E. H. (Ed.) (1990a). *Den store dialektboka [The large dialect book]*. Oslo: Novus.

Jahr, E. H. (1990b). Dialekter og dialektbruk i Norge [Dialects and dialect use in Norway]. See Jahr (1990a), pp. 7–28.

Jahr, E. H. and O. Skare (Eds.) (1996). *Nordnorske dialektar [North Norwegian dialects]*. Oslo: Novus forlag.

Jain, S. and B. C. Wallace (2019). Attention is not explanation. 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)*, Minneapolis, Minnesota, pp. 3543–3556. Association for Computational Linguistics.

Jha, A. and R. Mamidi (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using Twitter data. In *Proceedings of the Second Workshop on Natural Language Processing and Computational Social Science*, pp. 7–16.

Johannessen, J. B., K. Hagen, L. Håberg, S. Laake, Å. Søftelandog, and Ø. A. Vangsnes (2009). Transkripsjonsrettleiing for ScanDiaSyn [Transcription guidelines for ScanDiaSyn]. Unpublished manual. Retrieved from `http://tekstlab.uio.no/nota/scandiasyn/Transkripsjonsrettleiing%20for%20ScanDiaSyn.pdf` (October 28, 2020).

Johannessen, J. B., J. J. Priestley, K. Hagen, T. A. Åfarli, and Ø. A. Vangsnes (2009). The Nordic Dialect Corpus: An advanced research tool. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, Odense, Denmark, pp. 73–80. Northern European Association for Language Technology (NEALT). `http://tekstlab.uio.no/nota/scandiasyn/`.

Kåsen, A., K. Hagen, J. B. Johannessen, A. Nøklestad, and J. Priestley (2020). Comparing methods for measuring dialect similarity in Norwegian. In *Proceedings of the 12 Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, pp. 5343–5350. European Language Resources Association.

Kristoffersen, G. (2000). *The phonology of Norwegian*. The Phonology of the World's Languages. Oxford: Oxford University Press.

Kvale, K. and A. K. Foldvik (1997). Old customs die hard: The 'new' way of counting in Norwegian is still new. In *Proceedings of the 9th Swedish Phonetics Conference (Fonetik-97)*, Umeå, pp. 113–116.

Laake, S., I. F. Gjermundsen, A. Grov, K. Hagen, J. B. Johannessen, K. Kinn, A. Lykke, and E. Olsen (2011). Nordisk dialektkorpus: Oversettelse fra dialekt til bokmål [Nordic dialect corpus: Translation from dialect to Bokmål]. Unpublished manual. Retrieved from `http://tekstlab.uio.no/nota/scandiasyn/oversetter-veiledning.pdf` (October 28, 2020).

Le, H., L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab (2020). Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, Marseille, France, pp. 2479–2490. European Language Resources Association.

Lie, S. (1990). Oslo bymål [Oslo dialect]. See Jahr (1990a), pp. 179–184.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory 37*(1), 145–151.

Mæhlum, B. and U. Røyneland (2012). *Det norske dialektlandskapet: Innføring i studiet av dialekter [The Norwegian dialect landscape: Introduction to dialectology]*. Cappelen Damm Akademisk.

Malmasi, S. and M. Zampieri (2017). German dialect identification in interview transcriptions. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Valencia, Spain, pp. 164–169. Association for Computational Linguistics.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), *Advances*

in Neural Information Processing Systems 26 (NIPS 2013), Volume 26. Curran Associates, Inc.

Pamungkas, E. W., V. Basile, and V. Patti (2020). Misogyny detection in Twitter: A multilingual and cross-domain study. *Information Processing & Management 57*(6), 102360.

Papazian, E. (2008). De—dykk—dere. Andre person flertall i nordisk, særlig norsk [De—dykk—dere. Second person plural in Scandinavian languages, Norwegian in particular]. *Maal og Minne 100*(1), 69–97.

Papazian, E. and B. Helleland (2005). *Norsk talemål: lokal og sosial variasjon [Spoken Norwegian: local and social variation]*. Kristiansand: Høyskoleforlaget.

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

Poletto, F., V. Basile, M. Sanguinetti, C. Bosco, and V. Patti (2020). Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 1135–1144. Association for Computing Machinery.

Risch, J., R. Ruff, and R. Krestel (2020). Offensive language detection explained. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France, pp. 137–143. European Language Resources Association (ELRA).

Rodríguez-Sánchez, F., J. Carrillo-de Albornoz, and L. Plaza (2020). Automatic classification of sexism in social networks: An empirical study on Twitter data. *IEEE Access 8*, 219563–219576.

Sandøy, H. (1990). Vestlandet—der fjordane batt folket saman [West Norway—where the fjords bound people together]. See Jahr (1990a), pp. 63–87.

Sandøy, H. (1991). *Norsk dialektkunnskap [Norwegian dialect studies]* (2 ed.). Oslo: Novus.

Serrano, S. and N. A. Smith (2019). Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, pp. 2931–2951. Association for Computational Linguistics.

Sun, X. and W. Lu (2020). Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 3418–3428. Association for Computational Linguistics.

van Ommeren, R. and P. M. Kveen (2019). Det folkelingvistiske konseptet "tonefall": Ei sosiolingvistisk utforskning av prosodiens indeksikalitet [The folk linguistic concept of "accent/prosody": A sociolinguistic study of the indexicality of prosody]. *Maal og Minne 111*(1).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Wallace, B. C. (2019). Thoughts on "Attention is not not explanation". Blog post. Retrieved from `https://medium.com/@byron.wallace/thoughts-on-attention-is-not-not-explanation-b7799c4c3b24` (April 20, 2021).

Wiegreffe, S. and Y. Pinter (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 11–20. Association for Computational Linguistics.

Wieling, M. and J. Nerbonne (2011). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech & Language 25*(3), 700–715.

Yang, Z., D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, USA, pp. 1480–1489. Association for Computational Linguistics.

Zampieri, M., S. Malmasi, N. Ljubešić, P. Nakov, A. Ali, J. Tiedemann, Y. Scherrer, and N. Aepli (2017). Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017)*, Valencia, Spain, pp. 1–15. Association for Computational Linguistics.

Zampieri, M., S. Malmasi, P. Nakov, A. Ali, S. Shon, J. Glass, Y. Scherrer, T. Samardžić, N. Ljubešić, J. Tiedemann, C. van der Lee, S. Grondelaers, N. Oostdijk, D. Speelman, A. van den Bosch, R. Kumar, B. Lahiri, and M. 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 2018)*, Santa Fe, New Mexico, USA, pp. 1–17. Association for Computational Linguistics.

Zampieri, M., S. Malmasi, Y. Scherrer, T. Samardžić, F. Tyers, M. Silfverberg, N. Klyueva, T.-L. Pan, C.-R. Huang, R. T. Ionescu, A. M. Butnaru, and T. 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 2019)*, Ann Arbor, Michigan, USA, pp. 1–16. Association for Computational Linguistics.