Investigating and Explaining Feature and Representation Learning in Translationese Classification

Anonymous ACL submission

Abstract

Recent work has shown that neural feature- and representation-learning approaches, and specifically the BERT model, demonstrates superior performance over traditional manual feature en-005 gineering and an SVM classifier for the task of translationese classification for various source and target languages. However, to date it is un-007 clear whether the performance differences are due to better representations, better classifiers or both. Moreover, it remains unclear whether the features learnt by BERT overlap with com-011 monly used manual features. To answer these, 012 we exchange features between BERT-based and SVM classifiers, and show that, an SVM fed with BERT representations performs at the level of the best BERT classifiers, and BERT learning and using hand-crafted features per-017 forms at the level of traditional classifiers using hand-crafted features. Our experiments indi-019 cate that our hand-crafted feature set does not provide any additional information that BERT has not learnt already, and is likely to be a subset of features automatically learnt by BERT. Finally, we apply Integrated Gradients to examine token importance for the BERT model, and find that part of its top performance results are due to just topic differences and spurious 027 correlations with translationese.

1 Introduction

041

Translationese is a descriptive (non-negative) cover term for the systematic differences between translated and originally authored text in same language (Gellerstam, 1986). Some aspects of translationese such as source interference (Toury, 1980; Teich, 2003) are language dependent, others are presumed universal, e.g. simplification, explicitation, overadherence to target language linguistic norms (Volansky et al., 2015) in the products of translations. While translationese effects can be subtle, and even human experts may not be able to reliably distinguish between original texts and professional translations (Tirkkonen-Condit, 2002), corpus-based studies (Baker et al., 1993) and, in particular, machine-learning classifier based studies (Rabinovich and Wintner, 2015; Volansky et al., 2015; Rubino et al., 2016; Pylypenko et al., 2021) clearly reveal the differences. In this paper we focus on machine-learning classifier based research on translationese. Here, typically a classifier is trained to distinguish between original and translated texts (in the same language). Until recently, most of this research (Baroni and Bernardini, 2005; Volansky et al., 2015; Rubino et al., 2016) used manually defined, often linguistically inspired, feature-engineering based sets of features, (mostly) using support vector machines (SVM). Once a classifier is trained, feature importance and ranking methods are used to reason back to what aspects of the input is responsible for (i.e. explains) the classification. More recently, a small number of papers explored feature- and representation-learning neural network based approaches to translationese classification (Sominsky and Wintner, 2019). In a systematic study Pylypenko et al. (2021) shows that feature- and representation-learning deep neural network-based approaches (in particular BERTbased, but also other neural approaches) to translationese classification substantially outperform handcrafted feature-engineering based approaches using SVMs. However, to date, two important questions remain: (i) it is not clear whether the substantial performance differences are due to learned vs. handcrafted features, the classifiers, or the combination of both, and (ii) what the neural feature and representation learning approaches actually learn. The contributions of our paper are as follows:

043

044

045

046

047

050

051

052

053

057

058

059

060

061

062

063

064

065

066

067

068

069

070

071

073

074

075

077

078

079

081

 we address (i) by feeding BERT-based learned features to SVMs and by letting BERT architectures learn handcrafted features, as well as feeding the handcrafted features into BERT as embeddings. Our experiments show that SVMs using BERT-learned features perform on a par with our best BERT-

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

182

133

134

translationese classifiers. Moreover, BERT using handcrafted features only performs at most as good as the SVM classifier.

089

094

098

100

102

103

104

105

106

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

2. we present the first steps to address (ii)using attribution-based explainable AI approaches (XAI) on our best performing full feature- and representation-learning BERT model and on BERT models that are pretrained to predict handcrafted features and then fine-tuned for translationese classification. We present evidence that at least part of the high classification accuracy of BERT is due to names of places and countries, suggesting that part of the classification is topicand not translationese-based (source texts in Spanish translated to English e.g. may have a higher likelihood of talking about Spanish places). Moreover, some top features suggest that there might be certain spurious correlations within our dataset.

To the best of our knowledge this is the first paper that shows that feature- and representationlearning rather than the classifier is responsible for the substantial performance gap between deep neural networks and machine learning approaches using handcrafted features. It is also the first paper that uses XAI methods to (begin to) explain what neural methods learn in translationese classification.

Finally, translationese research is not just an "academic" exercise in basic research into aspects of how translation works, but an important research topic in machine translation evaluation (Stymne, 2017; Toral et al., 2018; Freitag et al., 2019; Edunov et al., 2020; Graham et al., 2020) and in further improving machine translation (Riley et al., 2020).

2 Related Work

(Kaas et al., 2020; Prakash and Tayyar Madabushi, 2020; Lim and Madabushi, 2020) combine BERTbased and handcrafted features in an ensemble manner in order to improve over BERT's accuracy, often by concatenating the pooled output of BERT with a handcrafted feature vector (sometimes additionally encoded by another network) and feeding them into another classifier. They show that even though BERT representations are powerful, carefully picked handcrafted features may still provide additional information that aids the task in hand. We also aim to verify if there is an overlap between the features learnt by BERT and our handcrafted feature set. In some cases the classifier used for ensembling BERT and handcrafted features is a Support Vector Machine (Kazameini et al., 2020; Ray and Garain, 2020).

Explainability methods for neural networks have not been widely explored for translationese classification. Since many previous works have used the traditional feature-engineering method, they often quantify handcrafted feature importance. Techniques used for that include looking at SVM feature weights (Avner et al., 2016; Pylypenko et al., 2021), correlation (Rubino et al., 2016), Information Gain (Ilisei et al., 2010), Chi-square (Ilisei et al., 2010), decision trees or random forests (Rubino et al., 2016; Ilisei et al., 2010), ablating features and observing the change in accuracy (Baroni and Bernardini, 2005; Ilisei et al., 2010), training separate classifiers on each individual feature (or feature set) and comparing accuracies (Volansky et al., 2015; Avner et al., 2016). For n-grams, then difference in frequencies between the original and translationese classes (Koppel and Ordan, 2011; van Halteren, 2008), and the contribution to the symmetrized Kullback-Leibler Divergence between the classes (Kurokawa et al., 2009) have been used.

As for looking into the neural network performance, Pylypenko et al. (2021) quantify whether hand-crafted features can explain the variance in the predictions of neural models, such as BERT, LSTM, and Simplified Transformer, by training per-feature linear regression models to output the predicted probabilities of the neural models and computing the R^2 measure. They find that most of the top features are either POS-perplexity-based, or bag-of-POS features. However, this method treats the neural network as a black-box, whereas we use a method that accesses the internals of the model.

In our work we use the Integrated Gradients method (Sundararajan et al., 2017) the method provides attribution scores for the input with respect to a certain class. It involves calculating the integral of gradients with respect to the input along the path from a certain baseline (in our case, PAD tokens) to the input.

3 Experimental Settings

3.1 Data

For our experiments, we use the monolingual German dataset in the Multilingual Parallel Direct Europarl (MPDE) (Amponsah-Kaakyire et al., 2021) corpus. The set contains 42k paragraphs with half of the texts German originals and the other half translations into German from Spanish. The average length is 80 tokens per training sample. Since there exists a problem with pivot translations in Europarl (Bogaert, 2011), the DE-ES dataset contains only data from before 2004, when the pivot system was introduced. We additionally use a heldout corpus of around 30k paragraphs for estimating language models and *n*-gram quartile distributions. This corpus consists of originally produced texts only. For the heldout corpus, we sample texts from Europarl proceedings from 2004 onwards, since original data is not affected by the pivot translation problem.

183

184

188

189

190

191

192

194

195

196

197

198

199

204

205

211

212

213

214

215

216



Figure 1: Mapping handcrafted features to embeddings (Section 3.4).

3.2 Base Setup

A fair comparison between learned and handcrafted features and two classifiers is non-trivial. We feed learned features into SVMs, and (i) let BERT learn handcrafted feature vectors used by SVMs and (ii) feed handcrafted feature vectors as embeddings into the BERT model. We compare this with full feature and representation learning settings with BERT, and SVMs with handcrafted features. To test this for translationese classification we reproduce the models from Pylypenko et al. (2021):

- a linear SVM with 108 handcrafted features (with surface, lexical, unigram bagof-PoS, language modelling and n-gram frequency distribution features), [handcr.features+SVM]
- a pretrained BERT-base model (12 layers, 768 hidden dimensions, 12 attention heads)

fine-tuned	on	translationese	classification.					
pretrained-BERT-ft]								

217

218

219

220

221

224

225

226

228

229

232

233

234

235

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

254

256

257

259

260

We use multilingual BERT (Devlin et al., 2019) (BERT-base-multilingual-uncased), and fine-tuning is done with the *simpletransformers*¹ library. We use a batch size of 32, learning rate of $4 \cdot 10^{-5}$, and the Adam optimiser with epsilon $1 \cdot 10^{-8}$. We estimate *n*-gram language models with SRILM (Stolcke, 2002) and do POS-tagging with SpaCy.²

3.3 SVM Classifier with BERT Features

We train a SVM with linear kernel on the features learnt by the pretrained BERT model fine-tuned on the translationese classification task. We use the output of the BERT pooler, which selects the last layer [CLS] token vector, with linear projection and tanh activation as our feature vector. We use:

- 1. BERT's 768-dim pooled vector output, [pretrained-BERT-ft-output+SVM]
- a 108-dim PCA projection of the vector. [pretrained-BERT-ft-output+PCA(108 dim) +SVM]

The later allows us to match the handcrafted feature vector dimensionality.

3.4 BERT with Handcrafted Features

Making neural networks use handcrafted features in our comparison is non-trivial. We design two strategies.

Pretraining on handcrafted feature prediction.

First, we train a BERT-base model from scratch, using it to predict the 108 dimension vectors representing handcrafted features originally used in training the SVM [**BERT-reg-full**]. The weights of this model encode the information of the hand-crafted features. With this pretrained model,

- we freeze the weights and train a classifier on top for translationese classification, [BERTr2c-full-frozen]
- 2. we do not freeze but fine-tune on the translationese classification task. [**BERT-r2c-full-ft**]

We explore the same with a smaller BERT model with only 6 layers instead of 12 [**BERT-reg-half**]. We then load the weights of the small 6 layer model into the embedding layer and the first 6 layers of a 12 layer non-pretrained BERT-base model and:

¹github.com/ThilinaRajapakse/ simpletransformers

²https://spacy.io/

Model	Test accuracy (%)
handcrfeatures + SVM	73.2±0.1
pretrained-BERT-ft-output + PCA(108 dim) + SVM	92.0±0.0
pretrained-BERT-ft-output + SVM	92.0±0.0
BERT-r2c-full-frozen-output + PCA(108 dim) + SVM	70.3±0.1
BERT-r2c-full-frozen-output + SVM	74.9±0.7
pretrained-BERT-ft	92.2±0.2
fromScratch-BERT	89.3±0.3
BERT-r2c-full-frozen	59.6±0.1
BERT-r2c-full-ft	89.3±0.4
BERT-r2c-half-frozen	67.5±0.4
BERT-r2c-half-ft	89.0±0.3
BERT-f2c $L = 1$	57.1±10.1
BERT-f2c $L = 80$	72.8±0.2
BERT-f2c $L = 256$	72.7±0.2
pretrained-BERT-f2c $L = 80$	68.0±2.1

Table 1: Translationese classification accuracy for all settings (average and standard deviation over 5 runs). All of the models were trained/fine-tuned for the translationese classification task.

3. we freeze the loaded weights in the first 6 layers and train the remaining 6 layers and classifier on the translationese classification task, [**BERT-r2c-half-frozen**]

261

262

267

269

272

273

274

275

276

4. we do not freeze but fine-tune it on the translationese classification task with randomlyinitialised weights for the other 6. [BERTr2c-half-ft]

Interestingly, according to the losses when training for predicting the handcrafted features, BERTreg-half performs comparably to the BERT-reg-full (0.0041136 vs 0.0041148).

We also train a BERT-base model with the same settings from scratch on the translationese classification task as a baseline for all BERT models. [fromScratch-BERT].

Mapping handcrafted features to embeddings. 277 Even though the very low MSE results indicate 278 that both versions of BERT-reg are able to learn 279 handcrafted features well, using them in terms of frozen layers in translationese classification leads 281 to low classification performance. This motivates us to explore an alternative way of encoding hand-283 crafted features: we convert the single vector of handcrafted features of dimension D (108 in our experiments) into a sequence of embeddings in BERT's layers format, that is, length of feature 287 embedding sequence L times the dimension of the hidden states H (768), while preserving the information of the single vector. (Fig. 1)

To do this, we consider a batch of tokens with size B and take in the handcrafted features as a $B \times D$ -dimensional input to the BERT model and generate feature embeddings by passing the features through 2 linear layers as follows. We first unsqueeze the $B \times D$ input to $B \times 1 \times D$ dimensions and reshape it as $B \times D \times 1$. This is passed to the first linear layer. The resulting $B \times D \times L$ dimensional output is reshaped as $B \times L \times D$ and fed as input to the second linear layer which outputs a $B \times L \times H$ -dimensional output as the feature embeddings. 291

292

293

295

296

297

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

This hand-crafted feature embedding replaces BERT's embedding layer and serves as input to the first BERT layer. The resulting BERT model is trained on the translationese classification task. We experiment with three different values for *L*: 1, 80 (average length of our training samples) and 256 (maximum input for BERT). [BERT-f2c L=1, BERT-f2c L=80, BERT-f2c L=256]

4 Translationese Classification

Table 1 summarises results of the different translationese classification settings. As for feeding pooled output of BERT into the SVM model, we can observe that the accuracy is a lot higher comparing to feeding handcrafted features, even when the BERT vector dimensionality is reduced to match the amount of handcrafted features. This emphasizes the fact that the features learnt by BERT are superior to our current set of manual features.

Legend	: 🗖 Negative	e 🗆 Neutral 🗖	Positive	
True Label	Predicted Label	Attribution Label	Attribution	pretrained-BERT-ft: Word Importance
Laber	Laber	Laber	Store	[CLS] das problem an der wur ##zel zu nack ##en und die gegenwart ##ige krise zu über ##wind ##en bedeutet, der menschlichen entwicklung
				vor ##rang ein ##zur ##au ##men wir mussen in zwei richtung ##en vor ##gehen kurz - und mittel ##fr ##isti ##g mussen wir den schwernunkt
1	1	1	2.76	auf die er ##nab ##rung ##ssi ##cher ##heit und die er ##nab ##rung ##sso ##uve ##rani ##tat lege ##n. zwei en ##scheiden ##de assecte
•	(1.00)	•	2	den beute existieren ##den hunger zu mild ##ern langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten
				lander gefordert werden [SEP]
				ander Beiordert werden - [PTF]
Legend	: 🗖 Negative	Neutral	Positive	
Label	Label	Label	Score	fromScratch-BERT: Word Importance
				[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung
				vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##en vor ##gehen . kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt
1	1	1	2.66	auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n, zwei ent ##scheiden ##de aspekte, um
	(0.97)			den heute existieren ##den hunger zu mild ##ern . langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten
				lander gefordert werden . [SEP]
Lange			Desition	
True	Predicted	Attribution	Attribution	
Label	Label	Label	Score	BER1-r2c-luil-it: Word Importance.
				[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung
				vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##en vor ##gehen . kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt
1	(0.89)	1	2.62	auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte, um
				den heute existieren ##den hunger zu mild ##ern . langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten
				lander gefordert werden . [SEP]
Logond	Nogetive	Neutral 🗖	Desitive	
Legend True	: Negative Predicted	e □ Neutral ■ Attribution	Positive Attribution	
Legend True Label	: ■ Negative Predicted Label	C □ Neutral ■ Attribution Label	Positive Attribution Score	BERT-r2c-full-frozen: Word Importance.
Legend True Label	: Negative Predicted Label	C Neutral ■ Attribution Label	Positive Attribution Score	BERT-r2c-full-frozen: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung
Legend True Label	Predicted	□ Neutral ■ Attribution Label	Positive Attribution Score	BERT+2c-full-frozen: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##en vor ##gehen . kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt
Legend True Label	I Negative Predicted Label	I Neutral Neutral Attribution	Positive Attribution Score -1.92	BERT-r2c-full-frozen: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##au ##men. wir mussen in zwei richtung ##en vor ##gehen. kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##si ##cher ##heit und die er ##nah ##rung ##ssi ##ave ##rani ##tal lege ##n "zwei ent ##scheiden ##de aspekk", um
Legend True Label	I (0.51)	I Neutral Attribution	Positive Attribution Score -1.92	BERT-r2c-full-frozen: Word Importance. [CLS] das problem an der wur ##zel zu pack men und die gegenwart ##ige kriss zu uber ##wind ##en bedeutet], der menschlichen entwicklung vor ##rang ein ##zur ##au ##nen . wir mussen in zwei richtung ##en vor ###gehen . kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi #echer ##heitu mid ein er #mah ##rung ##sta#wur ##mai ##tal Lege ##n, zwei ein ##scheiden ##de spekke zum den heute existieren ##den hunger zu mild ##em . langer ##fr ##isti ##g muss das nach ##halt ##fige wachst ##um fur die weniger entwickelten
Legend True Label	Regative Predicted Label I (0.51)	I Neutral Attribution	Positive Attribution Score -1.92	BERT-72-full-frozen: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##aa ##nen. Wir mussen in zwei richtung ##en vor ##gehen. kurz - und mittel ##f ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##si ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n "zwei ent ##scheiden ##de aspekte "um den heute existieren ##en imden ##en langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden. [SEP]
Legend True Label 1	Regative Predicted Label 1 (0.51)	 Neutral Attribution Label Neutral Neutral 	Positive Attribution Score -1.92 Positive	BERT-r2c-full-frozen: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##au ##nen. wir mussen in zwei richtung ##en vor #gehen. kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tal lege ##n "zwei en ##scheiden ##de aspekte "um den heute existieren ##den hunger zu mild ##en. langer ##fr ##isti ##g muss das nach ##halt ##tige wachst ##rum fur die weniger entwickelten lander gefordert werden. [SEP]
Legend True Label 1 Legend True	I Negative Predicted Label	Neutral Attribution Label Neutral Methods	Positive Attribution Score -1.92 Positive Attribution	BERT-r2c-full-frozen: Word Importance. [CLS] das problem an der wur ##zel zu pack men und die gegenwart ##ige krise zu ube ##wind ##en bedeutet "der menschlichen entwicklung tor ##rang ein ##zur ##au ##nen. wir mussen in zwei richtung ##en vor ##gehen. kurz - und mitel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##vor ##frail lege ##n_zwei ent ##scheiden ##de aspekte um den heute existieren ##den hunger zu mild ##ern. langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden. [SEP] BERT-r2c-half-ft: Word Importance.
Legend True Label 1 Legend True Label	I (0.51)	C Neutral Attribution Label I I Neutral Attribution Label	Positive Attribution Score -1.92 Positive Attribution Score	BERT-r2c-full-frozen: Word Importance. [CLS] das problem an der wur ##zel zu pack men und die gegenwart ##ige krise zu uber ##wind ##en bedeutet], der menschlichen entwicklung vor ##rang ein ##zur ##au ##nen . wir mussen in zwei richtung ##ar vor ##ram immediate ##fi ##isti ##g mussen wir den schwerpunkt auf die er ##nan ##rung ##sisi #wichet wiel er wirmshaft#rung ##sisi ##g musse das nach ##halt ##fige wachst ##our file vor ##echeiden ##de specket.com den heute existieren ##den hunger zu mild ##ern . langer ##fr ##isti ##g muss das nach ##halt ##fige wachst ##our file weniger entwickelten lander gefordert werden . [SEP] BERT-r2c-half-ft: Word Importance.
Legend True Label I Legend True Label	: Negative Predicted Label 1 (0.51) : Negative Predicted Label	C Neutral Attribution Label Neutral C Neutral Attribution Label	Positive Attribution Score -1.92 Positive Attribution Score	BERT-r2c-full-frzeen: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krisse zu uber ##wind ##en bedeutet "der menschlichen entwicklung wei ##rang ein ##zur ##au ##nen. wir mussen in zwei richtung ##en over ##gehen. kurz - und mittel ##fr ##stil ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tal lege ##n "zwei ent ##scheiden ##de aspekte "um den heute existieren ##den hunger zu mitd ##en. langer ##fr ##isti ##g muss das nach ##halt ##ing wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP] BERT-r2c-half-ft: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenward##fige krisse zu uber ##wind ##en bedeutet, der menschlichen entwicklung to part and and an geno ##en angen wir ##en und die gegenward##fige krisse zu uber ##wind ##en bedeutet, der menschlichen entwicklung to part and and and an geno ##en angen wir ##en und die gegenward##fige krisse zu uber ##wind ##en bedeutet, der menschlichen entwicklung to part and and and an geno ##en angen wir ##en bedeutet in der menschlichen entwicklung to part and and and an geno ##en angen ##en angen ##en bedeutet in der menschlichen entwicklung to part angen ##en angen ##en angen ##en bedeutet in der menschlichen entwicklung to part angen ##en angen ##en angen ##en angen ##en bedeutet in der menschlichen entwicklung to part angen ##en angen ##en angen ##en angen ##en bedeutet in der menschlichen entwicklung to part angen ##en angen ##en angen ##en angen ##en bedeutet in der menschlichen entwicklung to part angen ##en angen ##en angen ##en angen ##en angen ##en bedeutet in der menschlichen entwicklung to part angen ##en
Legend True Label 1 Legend True Label	Negative Predicted Label (0.51) Negative Predicted Label l	Neutral Attribution Label Neutral Neutral Attribution Label	Positive Attribution Score -1.92 Positive Attribution Score	BERT-22-full-frozen: Word Importance. [CLS] das problem an der wur ##zet zu maksen in zwei richtung ##con vor ##geken. kuzz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt und die er #mah ##rung ##ssi ##cher ##heit und die er #mah ##rung ##ssi ##uwe ##ran ##itallege ##n ; zwei en ##scheiden ##de aspekte ; um den heute existieren ##den hunger zu mild ##em. langer ##fr ##isti ##g muss das nach ##halt ##fig #wachst ##rum fur die weniger entwickelten lander gefordert werden . [SEF] BERT-22-half-ft: Word Importance. [(CLS] das problem an der wur ##zet zu pack ##en und die gegerwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##men. wir mussen in zwei richtung ##en vor ##geken kuzr - und mittel ##fr ##isti ##g mussen wir den schwerpunkt vor ##rang ein ##zur ##au ##men. wir mussen in zwei richtung ##en vor ##geken kuzr - und mittel ##fr ##isti ##g mussen wir den schwerpunkt vor ##rang ein ##zur ##au ##men. wir mussen in zwei richtung ##en vor ##geken kuzr - und mittel ##fr ##isti ##g mussen wir den schwerpunkt
Legend True Label 1 Legend True Label	Negative Predicted Label (0.51) Negative Predicted Label (0.96)	Neutral Attribution Label Neutral Attribution Label Attribution Label Attribution Label	Positive Attribution Score -1.92 Positive Attribution Score 3.05	BERT-r2c-full-frozen: Word Importance. [CLS] das problem an der wur ##zel u pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet] der menschlichen entwicklung vor ##rang ein ##zur ##au ##nen . wir mussen in zwei richtung ##en vor ##gehen. kurz - und nittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##sai #ehen . wir mussen in zwei richtung ##en vor ##gehen. kurz - und nittel ##fr ##isti ##g mussen wird en schwerpunkt auf die er ##nah ##rung ##sai #ehen . wir mussen in zwei richtung ##en vor ##gehen. kurz - und nittel ##fr ##isti ##g mussen wird en schwerpunkt BERT-r2c-half-fr: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##is vor ##gehen. kurz - und mittel ge#m. zwei ent ##schieden #de aspekke, um den heute existieren ##isti UCLS] das problem an der wur ##zel zu pack ##en und die gegenwart##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##isti ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##inen . wir mussen in zwei richtung ##isti ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##inen . wir mussen in zwei richtung ##isti ##ige krise zu uber ##wind ##en bedeutet , der menschlichen ##eistikung vor ##rang ein ##rau ##isti ##isti##isti ##isti ##isti ##isti ##isti ##isti
Legend True Label 1 Legend True Label	: Negative Predicted Label 1 (0.51) : Negative Predicted Label 1 (0.96)	Neutral Attribution Label Neutral Attribution Label Attribution Label Attribution Label	Positive Attribution Score -1.92 Positive Attribution Score 3.05	BERT-r2c-full-frzen: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krisse zu uber ##wind ##en bedeutet "der menschlichen entwicklung wei ##rang ein ##zur ##au ##nen. wir mussen in zwei richtung ##en over ##gehen. kurz - und mittel ##fr ##stil ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tal lege ##n. zwei ent ##scheiden ##de aspekte "um den heute existieren ##den hunger zu mitd ##en. langer ##fr ##isti ##g muss das nach ##halt ##fig wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP] EERT-r2c-half-ft: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenward###ige krisse zu uber ##wind ##en bedeutet, der menschlichen entwicklung vor ##rang ein ##zar ##au ##nen .wir mussen in zwei richtung ##en vor ##gehen. kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##en .wir mussen in zwei richtung ##en vor ##gehen. kurz - und mittel ##fr ##isti ## mussen in zwei ein #weichten und uie geken. wei men existieren ##den hunger zu mitd ##en .um ##isti ##g muss das nach ##halt ##ige #mit. zwei ent ##scheiden ##de schwerpunkt auf die er ##au ##au ##nen .wir mussen in zwei richtung ##en vor ##gehen. kurz - und mittel ##fr ##isti ## musseheiden ##de schwerpunkt auf die er ##au ##au ##au ##en .um #istie ##ge musse ##istie ##mittel ##fr ##istie ##istie ##istie ##for ##de schwerpunkt hunden betwe existieren ##den hunger zu mitd ##en .um ##istie ##ge musseheiden ##de schwerpunkt hunden betwe existieren ##den hunger zu mitd ##en .um ##istie ##mittel ##for ##istie ##i
Legend True Label 1 Legend True Label	Negative Predicted Label (0.51) Negative Predicted Label (0.96)	Neutral Attribution Label Neutral Attribution Label Attribution Label 1	Positive Attribution Score -1.92 Positive Attribution Score 3.05	BERT-2C-full-frozen: Word Importance. [CLS] das problem an der wur ##zet au pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung wor ##rang ein ##zur ##au ##nen. wir mussen in zwei richtung ##en vor ##gehen. kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##si ##cher ##heit und die er ##nah ##rung ##sis ##ave ##rani ##ial lege ##n _zwei ent ##scheiden ##de aspekke _um den heute existieren ##den hunger zu mild ##en _ langer ##fr ##isti ##g muss das nach ##halt ##frig ##sis ##ure ##arai ##ial [SLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##men. wir mussen in zwei richtung ##en vor ##geken. kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##sis ##cher ##heit und die er ##nah ##rung ##sis ##ure ##ani ##int ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##sis ##cher ##heit und die er ##nah ##rung ##sis ##ure ##ani ##int ##int ##int ##int ##int ##int ##int ##int [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##nen. wir mussen in zwei richtung ##sis ##au vor ##geken. kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##hah ##rung ##sis ##cher ##heit und die er ##nah ##rung ##sis ##ure ##au ##rai ##au lege ##n _ zwei ent ##scheiden ##de aspekke, um den houte existieren ##den hum mit ##fr ##isti ##g musse das nach ##halt ##fige wacht ##um uf die weniger entwickelten lander gefordert werden . [SEP]
Legend True Label 1 Legend True Label 1 Legend	Negative Predicted Label (0.51) Negative Predicted Label (0.96) Negative	Neutral Attribution Label Neutral Attribution Label Neutral Attribution Label Neutral	Positive Attribution Score -1.92 Positive Attribution Score 3.05 Positive	EERT-22-full-frozen: Word Importanc. [CLS] das problem an der wur ##zet zu pack ##en und die gegenwart ##ige krise zu ube ##wind ##en bedeutet, der menschlichen entwicklung vor ##rang ein ##zur ##au ##nen . wir mussen in zwei richtung ##en vor ##gehen. kurz - und nittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##sis ##ehet ##heit und die er ##nah ##rung ##sis ##uet ##an ##inge wachst ##uum für die weniger entwickelten lander gefordert werden . [SEP] [CLS] das problem an der wur ##zet zu pack ##en und die gegenwart##inge krise zu ube ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##en vor ##gehen. kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er #mah ##rung ##sis ##ehet ##heit und die er #mah ##rung ##sis ##uet und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er #mah ##rung ##sis ##ehet ##heit und die er #mah ##rung ##sis ##uet ##an ##inge wachst ##um für die weniger entwickelten lander gefordert werden . [SEP]
Legend True Label 1 Legend True Label 1	Negative Predicted Label (0.51) Negative Predicted Label (0.96) Negative Predicted Label Negative	Neutral Attribution Label Neutral Attribution Label Neutral Attribution Label Neutral Attribution Label	Positive Attribution Score -1.92 Positive Attribution Score 3.05 Positive Attribution Score	BERT-r2c-full-frozen: Word Importance. [(I.S.] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krisse zu ber ##wind ##en bedeutet],der menschlichen entwicklung wor ##rang ein ##zur ##au ##nen. Wir mussein izvei richtung ##en over ##en immediate ##fi ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##ure ##rani ##tat lege ##n _zwei ent ##scheiden ##de aspekte _um den heute existieren ##den hunger zu mild ##ern . langer ##fi ##isti ##g muss das nach ##halt ##fige wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP] [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart##ige kriss zu uher ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##au ##men . wir mussen in zwei richtung ##en vor ##gehen .kurz - und mittel ##fi ##isti ##g musses das nach die die er ##nah ##men #wirmung ##ssi #enen .wir mussen in zwei richtung ##en vor ##gehen .kurz - und mittel ##fi ##isti ##en scheiden ##de aspekte, um den heute existieren ##den hunger zu mild ##ern .langer ##fi ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden .[SEP] BERT-r2c-half-frozen: Word Importance.
Legend True Label I Legend True Label Legend True Label	Negative Predicted Label (0.51) Predicted Label (0.96) Negative Predicted Label Negative Predicted Label Negative	Neutral Attribution Label Neutral Attribution Label Neutral Attribution Label Attribution Label	Positive -1.92 Positive Attribution Score 3.05 Positive Attribution Score	EERT+2c-full-frzen: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet]der menschlichen entwicklung wer ##rang ein ##zu ##mac. Wir mussen in zwei richtung ##en vor ##gehen. kurz - und mitcle ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##nem. Wir mussen in zwei richtung ##en vor ##gehen. kurz - und mitcle ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##nem. Wir mussen in zwei richtung ##en vor ##geken. kurz - und mitcle ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rang ##si ##en. [SLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##men. wir mussen in zwei richtung ##en vor ##geken. kurz - und mitcle ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rang ##sis ##en. [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##nen. wir mussen in zwei richtung ##sis ##ge nusse das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander refordert werden. [SEP] EERT-2c-half-frzen: Word Importance. [CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden. [SEP] EERT-2c-half-frzen: Word Importance. [CLS] das problem an der wur ##zel zu mude ##en.
Legend True Label 1 Legend True Label Label	Negative Predicted Label (0.51) Negative Predicted Label (0.96) Negative Predicted Label Label	Neutral Attribution Label Neutral Attribution Label Neutral Constraint Attribution Label Attribution Label	Positive -1.92 Positive 3.05 Positive Attribution Score	EERT-22-full-frozen: Word Importanc. [CLS] das problem an der wur ##zet zu pack ##en und die gegenwart ##ige krise zu ube ##wind ##en bedeutet], der menschlichen entwicklung vor ##rang ein ##zur ##au ##nen. wir mussen in zwei richtung ##en vor ##gehen. kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er #mah ##rung ##sis #eher #theit und die er #mah ##rung ##sis ##uer. EERT-22-half-ft: Word Importance. [CLS] das problem an der wur ##zet zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##en vor ##gehen. kurz - und mittel ##fr ##isti ##g musses wir den schwerpunkt auf die er #mah #mrung ##sis #eher #theit und die er #mah ##rung ##sis ##uer werden. [SLS] das problem an der wur ##zet zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##run ##sis #eher #theit und die er #mah ##rung ##sis ##uer werden. kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er #mah #mrung ##sis #eher #theit und die er #mah ##rung ##sis #werden. [SLS] das problem an der wur ##zet zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##sen wer ##rai ##ige was das mach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden. [SEP] EERT-22-half-frozen: Word Importance. [CLS] das problem an der wur ##zet zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung wor ##rang ##ige ##ige ##ige ##ige krise zu uber ##wind ##in bedeutet , der menschlichen entwickleng lander gefordert werden. [SEP]
Legend True Label I Legend True Label I Legend True Label	Negative Predicted Label (0.51) Negative Predicted Label (0.96) Negative Predicted Label 1 (0.96) 1	Neutral Attribution Label Neutral Neutral Attribution Label Neutral Attribution Label Attribution Label Attribution Label	Positive Attribution Score -1.92 Positive Attribution Score 3.05 Positive Attribution Score 0.76	ERT-r2c-fulf-frzen: Word Importance. [CLS] das problem an der wur #rzel zu pack #ren und die gegenwart #rige krise zu ber #rwind #ren bedeutet], der menschlichen entwicklung word #rang ein #rzur #Bau #renne. wir mussein in zwei richtung #ren vor #ragehen. kurz - und mutiet #rfr #risti #rg mussen wir den schwerpunkt auf die er #mah #rinne, #rissein in zwei richtung #ren vor #ragehen. kurz - und mutiet #rfr #risti #rg mussen wir den schwerpunkt auf die er #mah #rinne, #rissein in zwei richtung #ren vor #ragehen. kurz - und mutiet #rfr #risti #rg mussen wir den schwerpunkt auf die er #mah #rinne #rissein #rweine #rissi #rene mutiet #rfr #rissi #rene mutiet #rfr #rissi #rene vertification #rene #rissi #rene vertification #rene vertification #rene vertification #rene vertification #rene vertification #rene vertification #rene vertification #rene vertification #rene vertification #rene verification #rene verification #r
Legend True Label 1 Legend True Label 1 Legend Label	 Negative Predicted Label (0.51) Negative Predicted Label (0.96) Negative Predicted Label (0.63) 	Neutral Attribution Label Neutral Attribution Label Neutral Attribution Label 1 1 1 1 1 1 1 1 1 1 1 1 1	Positive -1.92 Positive 3.05 Positive Attribution Score 3.05 Attribution Score 0.76	ERT-r2c-ful-frzen: Word Importance. (CLS) das problem an der wur ##rzet zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet], der menschlichen entwicklung wor ##rang ein ##zur ##au #men. wir mussen in zwei richtung ##aco ##aci ##

Figure 2: Layer Integrated Gradient saliency maps of input tokens contributing to the ground truth translationese label (here: translation). Comparison of different models.

When BERT is trained from scratch, translationese classification accuracy reduces by 2.93 percentage points, compared to the pretrained-BERTft. This suggests that pretraining on large data helps to encode additional information that turns out to be helpful in the translationese classification task.

321

322

323

324

325

328

329

331

333

334

335

337

339

341

342

As for BERT pretrained to predict the handcrafted features and frozen (BERT-r2c-full-frozen), it is assumed that BERT has learnt to encode the handcrafted features during pretraining. Nevertheless, its accuracy, albeit higher than a random guess, is lower by ~ 13 percentage points comparing to the SVM classifier. We perform an additional experiment, in order to check whether the difference in accuracy is due to BERT failing to sufficiently encode the handcrafted features during pretraining, or due to the SVM classifier being superior to the linear classification head of the BERT model. Namely, we train the SVM classifier on the pooled output of BERT-r2c-full-frozen model. The accuracy is around 75% (for both original and PCA-reduced dimensionality) which is as high as using SVM on handcrafted feature vectors. We conclude that

BERT encodes the handcrafted features sufficiently enough, but the linear classifier performs worse than an SVM in these conditions, possibly due to non-exhaustive hyperparameter search.

344

345

346

347

348

349

351

352

353

354

356

357

358

359

360

361

362

363

364

365

Further fine-tuning BERT, fully pretrained for handcrafted feature prediction (BERT-r2c-full-ft), for translationese classification results in accuracy comparable to BERT that was not pretrained on this task (fromScratch-BERT). This could suggest that our handcrafted feature set is either a subset of features learned by fromScratch-BERT, or that the handcrafted features are discarded during finetuning. The model where only the first 6 layers were pretrained (BERT-r2c-half-ft), achieves similar accuracy, likely due to the same reasons.

By contrast, freezing the 6 handcrafted feature prediction pretrained layers (BERT-r2c-halffrozen) largely reduces the accuracy, because the model only has access to the 6th layer embeddings that supposedly encode only the information about the handcrafted features, and does not have ability to extract its own features from the input text, due to its inability to tune the embeddings.

Legend True	I: Megative Predicted	Neutral Attribution	Positive Attribution	
Label	Label	Label	Score	pretrained-BER1-ft: word importance
0	0 (1.00)	0	4.82	[CLS] j\u00e9habe vers ##tandnis dafur, dass ein breite ##r.Kom##sense Somithema umwelt ##haft ##ung nicht immer moglich ist, marche date ##um mochte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ## marche det ##um nen der recht ##ums ##schuss hat eine klare ; ma ### ##hme des bericht ##s des aus ##achuse ##es fur recht und binnen ##marche det ##um großen schritt in richtung eines verb ##esse ##rten umwelt ##schut ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs ##timmung nachi ##druck ##lich unterstutzt werden. [SEP]
Legend True Label	I: Megative Predicted Label	□ Neutral ■ Attribution Label	Positive Attribution Score	fromScratch-BERT: Word Importance
0	0 (1.00)	0	1.41	[CLS] ich habe vers ##tandnis dafür, dass ein brüte/##t on ##sens beim thema umwelt ##haft ##ung nicht immer moglich ist, gerade dat ##um mochte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht um binnen ##markt wer ##ben, denn der recht ##us ##schuss hat eine klare, ma ##B ##vlole und pra ##kikk ##able stellung ##nahme ab #Begeeben. der bericht ##st ellt einen großen schritt in richtung eines verb ##esse ##net unwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs ##timmung nach ##druck ##lich unterstutzt werden. [SEP]
Legend True Label	l: Mcgative Predicted Label	□ Ncutral ■ Attribution Label	Positive Attribution Score	BERT-r2c-full-ft: Word Importance.
0	0 (1.00)	0	0.89	[CLS] ich habe vers ##tundnis datur, dass ein breite ##r kon ##sens beim thema umwelt ##haft ##ung nicht immer moglich ist]gerade dar ##um mechte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ##markt wer ##ben , denn der rechts ##aus ##schuss hat eine kare, ma ##B ##volle und par ##ktik ##able stellung ##nahme ab ##gegeben. der bericht stellt einen großen schrift in fichtung eines verb #=esse ##iten unwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##n morgen bei der abs ##timmung nach ##druck ##lich unterstutzt werden. [SEP]
Legend True Label	I: Negative Predicted Label	Neutral Attribution Label	Positive Attribution Score	BERT-r2c-full-frozen: Word Importance.
0	1 (0.51)	0	0.16	[CLS] ich habe vers ##tandnis dafur; dass ein breite ##r kon ##sens beim thema unwelt ##haft ##ung nicht immer moglich ist. gerade dar ##um mochte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ##markt wer ##ben i denn der rechts ##taus ##schuss hat eine klare ; ma ##B ##volle und pn ##kik ##able stellung ##nahme ab ##gegeben. der bericht stellt einen großen schritt in richtung eines verb ##uses ##rten unwelt ##schuz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs ##timmung nach ##druck ##lich unterstutzt werden. [SEP]
Legend	: Negative	Neutral	Positive	
True	Predicted	Attribution	Attribution	BERT-r2c-half-ft: Word Importance.
Label 0	Label 0 (1.00)	Label	5core	[CLS] ich habe vers ##tandnis dafur "idass ein breite ##r kon ##sens beim thema unwelt ##haft ##ung nicht immer moglich ist "igerade dar ##um mochte ich aber heute auch ganz besonders für die anna ##hme des bericht ##s des aus ##schuss ##es für recht und binnen ##markt wer ##ben denn der rechts ##aus ##schuss hat eine klare, ma ##ß ##volle und pra ##klik ##able stellung ##nahme ab ##gegeben "der bericht stellt einen großen schrift in richtung eines verb ##esse ##ten unwert ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs ##immung and ##dnuck ##lich unterstutzt werden. [SEP]
Legend True Label	l: Negative Predicted Label	Neutral Attribution	Positive Attribution Score	BERT-r2c-half-frozen: Word Importance.
0	0 (0.67)	0	0.98	[CLS] ich habe vers ##tandnis dafur, dass ein breite ##r kon ##sens beim thema unwelt ##haft ##ung nicht immer moglich ist gerade dar ##um mochte ich aber heute much ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ##markt wer ##ben , denn der rechts ##nas ##schuss hat eine klare , ma ##B ##volle und pna ##kik ##able stellung ##nahme ab ##gegeben. der bericht stellt einen großen schritt in richtung eines verb ##esse ##rten umwelt ##schuz ##es dar und sollte deshalb auch und] #immungen abe kärnds #ilbit huntersturt werden. [EPP]

Figure 3: Layer Integrated Gradient saliency maps of input tokens contributing to the ground truth translationese label (here: original). Comparison of different models.

The results of BERT-f2c models show that BERT, when fed the handcrafted features in the form of embeddings, can reach at most the same accuracy as the hancdr.-features+SVM approach, which suggests that the BERT architecture has no advantage over the SVM classifier in utilizing the handcrafted features for classification.

5 Layer Integrated Gradients Saliency

We compare input attributions of the ground truth classification label amongst pretrained-BERT-ft, fromScratch-BERT and the different settings of the translationese classification models pretrained on the hand-crafted feature prediction task. We use Layer Integrated Gradients from the Captum library (Kokhlikyan et al., 2020), and calculate the salience score for each token by averaging the attributions over the embedding dimension.

5.1 Comparing Models

367

371

373

374

375

376

377

381

384

385

Figure 2 displays Integrated Gradient attributions for a **translated** paragraph across different BERT models. Figure 3 shows attributions for an **original** paragraph.

Comparing the attributions of classification labels to sample inputs amongst the various settings of BERT, we observe that attributions are similar for fromScratch-BERT and the fine-tuned models: BERT-r2c-full-ft and BERT-r2c-half-ft. By contrast, freezing the weights in BERT-r2c-fullfrozen and BERT-r2c-half-frozen resulted in very different attributions from the fromScratch-BERT. For BERT-r2c-half-frozen the attributions are more peaked than for other models, with only a few tokens receiving large scores, and most tokens having scores close to zero. Notably, pretrained-BERTft displays a pattern that is overall similar to the BERT trained from scratch, but some attributions are reversed, and the peaks are on different tokens. 388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

For BERT-r2c-full-frozen, it can be seen that a substantial number of tokens with negative attributions have positive attributions in the model trained from scratch and also the fine-tuned models. However some attributions overlap, which suggests that BERT might be using some of the handcrafted features.

	BERT-1	r2c-full-ft	pretrained-BERT-ft			
Rank	Token	Avg attribution score	Rank	Token	Avg attribution score	
1	sagte	0.60	1	entstand	0.70	
2	gebiet	0.46	2	virus	0.63	
3	##dies	0.44	3	inti	0.60	
4	ansicht	0.43	4	sagte	0.58	
5	bezug	0.42	5	entdeckte	0.57	
6	neige	0.40	6	gras	0.57	
7	amt	0.40	7	nuts	0.56	
8	pre	0.40	8	nicaragua	0.55	
9	spanien	0.39	9	rekord	0.53	
10	sprechen	0.38	10	bilbao	0.53	
11	nuts	0.36	11	verfugte	0.53	
12	barcelona	0.34	12	bol	0.51	
13	;	0.33	13	colombia	0.51	
14	##bien	0.32	14	nis	0.51	
15	spanischen	0.32	15	och	0.49	
16	wiederholt	0.31	16	vorkommen	0.49	
17	einige	0.30	17	oecd	0.49	
18	##sprache	0.29	18	;	0.46	
19	weder	0.29	19	erklarte	0.45	
20	territorium	0.28	20	clinton	0.45	

Table 2: Top-20 tokens with highest average attribution score towards the **translationese** class in the test set. BERT-r2c-full-ft and pretrained-BERT.

5.2 Comparing Checkpoints

411

412

413

414

415

416

417

418

419

420

421

422

423

494

425

426

427

428

429

430

431

432

433

434

435

436

In Appendix A we provide additional results on examining training checkpoints for fromScratch-BERT and BERT-r2c-full-ft for an original and a translated paragraph.

Results indicate that for fromScratch-BERT some attributions change into opposite during training, whereas for BERT-r2c-full-ft the pattern appears to be already settled from the early checkpoints onwards, and does not change much over the course of fine-tuning. This may support the hypothesis that the handcrafted features are a subset of features learnt by BERT, and thus provide a favorable initialization of weights for fine-tuning for translationese classification.

5.3 Highest Average Attribution

In order to make the interpretation less local, we compute the top tokens with highest attribution on average across the test set. The results for each class for best-performing models (pretrained BERT and BERT-r2c-full-ft) are given in Tables 2 and 3.

For German translationese data translated from Spanish, some top tokens correspond to the geographical areas, where Spanish is spoken, e.g. "spanien", "barcelona", "spanischen" for BERTr2c-full-ft; "nicaragua", "colombia", "bilbao" for pretrained BERT. Likewise for original German data, some of the top tokens are German geographical names, e.g. "stuttgart" for pretrained BERT. The subword "##wald" also appears to be a common German toponymic suffix. This suggests that topic is one of the spurious clues that is used by BERT to determine the correct translationese class. This is also supported by the fact that some nouns that likely correspond to certain recurring discussion topics for only one class within our data sample, receive high attribution, e.g. "virus", "soja", "clinton", "orange" etc. The "ez" token, salient for the original class, appears to be a starting subword unit of the *EZB* abbreviation (Europäische Zentralbank). 437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

The "•" token having a high attribution for the class *originals* for both models might suggest a spurious correlation within the dataset, that is apparently utilized by BERT. The ";" token is deemed important for the translationese class by both models, which might also be a spurious correlation. Conversely, this could be an indication that clauses in Spanish are more often joint with the semi-colon, than in German, which was preserved in the translation.

For both models the Präteritum forms "sagte", "erklärte" etc. are also among the top tokens impor-

	BERT-r	2c-full-ft	pretrained-BERT-ft			
Rank	Token	Avg attribution score	Rank	Token	Avg attribution score	
1	##wegen	0.61	1	situations	0.37	
2	•	0.55	2	•	0.36	
3	eu	0.49	3	ria	0.34	
4	daraufhin	0.49	4	##lk	0.33	
5	finde	0.45	5	##iet	0.32	
6	##vo	0.45	6	golden	0.32	
7	gerne	0.43	7	sak	0.30	
8	##abb	0.42	8	turm	0.30	
9	##hrte	0.42	9	##emen	0.27	
10	ausbau	0.42	10	orange	0.27	
11	!	0.42	11	hang	0.26	
12	bekommen	0.42	12	##wald	0.25	
13	trips	0.41	13	1732	0.25	
14	ez	0.41	14	dobe	0.24	
15	##gemeinde	0.40	15	##pas	0.23	
16	vot	0.36	16	profits	0.22	
17	won	0.36	17	stuttgart	0.22	
18	geplant	0.35	18	soja	0.21	
19	demnach	0.35	19	r	0.21	
20	ja	0.35	20	ruth	0.21	

Table 3: Top-20 tokens with highest average attribution score towards the **original** class in the test set. BERT-r2c-full-ft and pretrained-BERT.

tant for recognizing translationese. One possible
explanation could be that the Perfekt form ("hat gesagt") is more common in German spoken language, and Präteritum is more common in writing. Therefore the translators, while translating Spanish
speeches into German, could have preferred to use the Präteritum form more common for writing.

6 Conclusion

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

485

486

487

488

489

This paper addresses two open questions in classification-based translationese research: (1) are the substantial performance differences between feature- and representation-learning and classical handcrafted feature based approaches due to (i) the difference in the features, (ii) the classifiers, or (iii) both, and (2) what do feature- and representationlearning based approaches actually learn?

We address (1) by exchanging features from both models examining a broad variety of settings. We confirm that SVMs perform as good as BERT when fed with features learnt by BERT. Likewise, BERT performs at the level of traditional SVM-based classification with handcrafted features SVMs, when fed with handcrafted features only. Our findings reveal that while pretraining on huge amount of data improves the classification accuracy, pretraining on handcrafted features does not guarantee an improvement on classification accuracy with respect to training from scratch.

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

509

510

511

512

513

514

To address question (2), we examine BERT's input attributions using Integrated Gradients Saliency for various settings and observe that attributions are indeed similar for the model trained from scratch and the fine-tuned models that were pretrained on handcrafted feature prediction.

Finally, analysis of top activated tokens in the test set suggests that at least part of BERT's strong translationese classification accuracy is based on topical differences between the classes (rather than "proper" translationese phenomena), the topical differences between the classes, and spurious correlations. The next step would be to control these factors, for instance by using named entity masking and cleaning/normalizing the corpus, in order to investigate whether BERT would still outperform the traditional approach under such conditions.

References

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2021. Do not rely on relay translations: Multilingual parallel direct Europarl. In *Proceedings for the First Workshop on Modelling Translation: Translatology*

- 515 516 517 518 521 526 527 528 531 532 533 534 535 539 540 541 542 543 545 546 547 548 549 551 552 553 554 556 564 567

571

- 560

519 520

in the Digital Age, pages 1-7, online. Association for

Ehud Alexander Avner, Noam Ordan, and Shuly Wint-

Mona Baker, Gill Francis, and Elena Tognini-Bonelli.

1993. Corpus linguistics and translation studies: Im-

plications and applications. In Text and Technology:

In Honour of John Sinclair, page 233–, Netherlands.

Marco Baroni and Silvia Bernardini. 2005. A New

Caroline Bogaert. 2011. Is absolute multilingualism

maintainable? The language policy of the Euro-

pean Parliament and the threat of English as a lingua

franca. Master's thesis, UGent. Faculteit Letteren en

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and

Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language under-

standing. In Proceedings of the 2019 Conference of

the North American Chapter of the Association for

Computational Linguistics: Human Language Tech-

nologies, Volume 1 (Long and Short Papers), pages

4171-4186, Minneapolis, Minnesota. Association for

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and

Michael Auli. 2020. On the evaluation of machine

translation systems trained with back-translation. In

Proceedings of the 58th Annual Meeting of the Asso-

ciation for Computational Linguistics, pages 2836-

2846, Online. Association for Computational Lin-

Markus Freitag, Isaac Caswell, and Scott Roy. 2019.

APE at scale and its implications on MT evaluation

biases. In Proceedings of the Fourth Conference on

Machine Translation (Volume 1: Research Papers), pages 34-44, Florence, Italy. Association for Com-

Martin Gellerstam. 1986. Translationese in Swedish

Yvette Graham, Barry Haddow, and Philipp Koehn.

2020. Statistical power and translationese in machine

translation evaluation. In Proceedings of the 2020

Conference on Empirical Methods in Natural Lan-

guage Processing (EMNLP), pages 72-81, Online.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and

Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In Computational Lin-

guistics and Intelligent Text Processing, pages 503-

511, Berlin, Heidelberg. Springer Berlin Heidelberg.

Association for Computational Linguistics.

novels translated from English. Translation studies

Approach to the Study of Translationese: Machine-

learning the Difference between Original and Trans-

Literary and Linguistic Computing,

John Benjamins Publishing Company.

ner. 2016. Identifying translationese at the word and

sub-word level. Digit. Scholarsh. Humanit., 31:30-

Computational Linguistics.

54.

lated Text.

21(3):259–274.

Wijsbegeerte.

guistics.

Computational Linguistics.

putational Linguistics.

in Scandinavia, 1:88-95.

Anders Kaas, Viktor Torp Thomsen, and Barbara Plank.

2020. Team DiSaster at SemEval-2020 task 11: Com-

bining BERT and hand-crafted features for identify-

ing propaganda techniques in news. In Proceedings

of the Fourteenth Workshop on Semantic Evaluation,

pages 1817–1822, Barcelona (online). International

Amirmohammad Kazameini, Samin Fatehi, Yash Mehta,

Sauleh Eetemadi, and Erik Cambria. 2020. Person-

ality trait detection using bagged SVM over BERT

word embedding ensembles. CoRR, abs/2010.01309.

Edward Wang, Bilal Alsallakh, Jonathan Reynolds,

Alexander Melnikov, Natalia Kliushkina, Carlos

Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020.

Captum: A unified and generic model interpretability

Moshe Koppel and Noam Ordan. 2011. Translationese

and its dialects. In Proceedings of the 49th An-

nual Meeting of the Association for Computational

Linguistics: Human Language Technologies, pages 1318–1326, Portland, Oregon, USA. Association for

David Kurokawa, Cyril Goutte, and P. Isabelle. 2009.

Automatic detection of translated text and its impact

on machine translation. In Proceedings of the 52nd

Annual Meeting of the Association for Computational

Wah Meng Lim and Harish Tayyar Madabushi. 2020.

corpus level information. CoRR, abs/2008.08547.

Anushka Prakash and Harish Tayyar Madabushi. 2020.

Incorporating count-based features into pre-trained

models for improved stance detection. In Proceed-

ings of the 3rd NLP4IF Workshop on NLP for Internet

Freedom: Censorship, Disinformation, and Propa-

ganda, pages 22-32, Barcelona, Spain (Online). In-

ternational Committee on Computational Linguistics

Daria Pylypenko, Kwabena Amponsah-Kaakyire, Koel

Dutta Chowdhury, Josef van Genabith, and Cristina

España-Bonet. 2021. Comparing feature-engineering

and feature-learning approaches for multilingual

translationese classification. In Proceedings of the

2021 Conference on Empirical Methods in Natural

Language Processing, pages 8596–8611, Online and

Punta Cana, Dominican Republic. Association for

Ella Rabinovich and Shuly Wintner. 2015. Unsuper-

Biswarup Ray and Avishek Garain. 2020. Factuality

tor machines. In *IberLEF@SEPLN*.

classification using bert embeddings and support vec-

vised identification of translationese. Transactions of

the Association for Computational Linguistics, 3:419-

Computational Linguistics.

Uob at semeval-2020 task 12: Boosting BERT with

library for pytorch.

(ICCL).

432.

9

Computational Linguistics.

Linguistics, Baltimore, Maryland.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin,

Committee for Computational Linguistics.

572

573

574

575

576

577

578

579

580

581

582

583

584

586

587

588

589

593

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

524

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7737–7746, Online. Association for Computational Linguistics.

627

637

643

647

651

652

653

654 655

656

663

674

675

676

677

678

679

- Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 960– 970, San Diego, California. Association for Computational Linguistics.
- Ilia Sominsky and Shuly Wintner. 2019. Automatic detection of translation direction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1131–1140, Varna, Bulgaria. INCOMA Ltd.
- Andreas Stolcke. 2002. SRILM An extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904.
- Sara Stymne. 2017. The effect of translationese on tuning for statistical machine translation. In *The 21st Nordic Conference on Computational Linguis-tics*, pages 241–246.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.
- Elke Teich. 2003. Cross-Linguistic Variation in System und Text. A Methodology for the Investigation of Translations and Comparable Texts. Mouton de Gruyter, Berlin.
- Sonja Tirkkonen-Condit. 2002. Translationese a myth or an empirical fact?: A study into the linguistic identifiability of translated language. *Target. International Journal of Translation Studies*, 14(2):207–220.
- Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way.
 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.
- Gideon Toury. 1980. *In Search of a Theory of Translation*. The Porter Institute for Poetics and Semiotics, Tel Aviv University, Tel Aviv.
- Hans van Halteren. 2008. Source language markers in EUROPARL translations. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 937–944, Manchester, UK. Coling 2008 Organizing Committee.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

10

A Appendix

686

687

688

690 691

692

693

696

697

698

699

700

701

702

A.1 Extra Information on MPDE dataset

We use version 2.0.0 of the MPDE dataset licensed under CC-BY 4.0. Specifically we use the *mono_de_es* train/dev/test splits of the German-Spanish language pair. Table 4 contains summary statistics of the data.

Split	Number of Examples
Train set	29580
Validation set	6366
Test	6344

Table 4: Dataset statistics

A.2 Extra Information on BERT models

With the exception of pretrained-BERT-ft, we use the *transformers* library.³ Training is done across 4 NVIDIA GeForce GTX TITAN X GPUs with a batch size of 8 per GPU. We use a learning rate of $3 \cdot 10^{-5}$ and train or fine-tune for 5 epochs. Table 5 shows the number of parameters of the different BERT variants. Parameter counts include the embedding and respective prediction (classifier or regression) layers.

Model	Num. Params (M)
fromScratch-BERT	177.85
BERT-reg-full	177.94
BERT-reg-half	135.41
BERT-r2c-*	177.85
BERT-f2c $L = 1$	177.46
BERT-f2c $L = 80$	177.52
BERT-f2c $L = 256$	177.66
pretrained-BERT-f2c $L = 80$	177.52

Table 5: Number of parameters of the various BERT models

³https://huggingface.co/transformers/ model_doc/bert.html

Legend True	Predicted Label	C Neutral ■ Attribution Label	Positive Attribution	fromScratch-BERT: Word Importance, checkpoint 100
Laber	Laber	Laber	Score	[CLS] ich habe vers ##tandnis dafur dass ein breite ##r kon ##sens beim thema umwelt ##haft ##ung nicht immer moglich ist gerade dar ##um
				mochte ich aber heute auch ganz besonders für die anna ##hme des bericht ##s des aus ##schuss ##es für recht und binnen ##markt wer ##ben
0	1	0	-1.54	denn der rechts ##aus ##schuss hat eine klare .ma ##ß ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben . der bericht stellt einen
	(0.54)			großen schritt in richtung eines verb ##ese ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs
				##timmung nach ##druck ##lich unterstutzt werden . [SEP]
egend	• 🗖 Negative	Neutral 🗖	Positive	
True	Predicted	Attribution	Attribution	fromScratch-BERT: Word Importance, checkpoint 200
Label	Label	Label	Score	[CI S] ich habe vers ##tandnis dafur dass ein breite ##r kan ##sens beim thema umwelt ##haft ##ung nicht immer modich ist gerade dar ##um
				nochte ich aber heute auch ganz besonders für die anna ##hme des bericht ##s des aus ##schuss ##es für recht und hinnen ##markt wer ##hen
0	1	0	-0.95	denn der rechts ##aus ##schuss hat eine klare ma ##B ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben der bericht stellt einen
	(0.55)			großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs
				##timmung nach ##druck ##lich unterstutzt werden . [SEP]
legend	: Negative	neutral	Positive	
True	Predicted	Attribution	Attribution	fromScratch-BERT: Word Importance, checkpoint 300
Label	Label	Label	Score	[CI S] ich hoha vare ##tandnie defue dass ain braita ##e kan ##sans haim thams umwalt ##hoft ##ung nicht immar maglich ist garada der ##um
				mochte ich aber heute auch ganz besonders für die anna ##hme des bericht ##s des aus ##schuss ##es für recht und binnen ##markt wer ##hen
0	0	0	0.17	denn der rechts ##aus ##schuss hat eine klare _ma ##ß ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben _der bericht stellt einen
	(0.64)			großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs
				##timmung nach ##druck ##lich unterstutzt werden . [SEP]
	No anti-	D Manufacel 🗖	Desition	
True	Predicted	Attribution	Attribution	fromScratch.BEPT: Word Importance checknoint 400
Label	Label	Label	Score	ron of the hear way have been been to be the way have been a first one of the second state of the second s
				[CLS] ich nabe vers ##tandnis darur, dass ein breite ##r kon ##sens beim thema umweit ##natt ##ung nicht immer möglich ist, geräde dar ##um mochte ich aber beute auch ganz besonders für die anna ##bme des bericht ##s des aus ##schuss ##se für recht und hinnen ##markt wer ##ben
0	0	0	0.77	denn der rechts ##aus ##schuss hat eine klare ma ##R ##volle und nra ##ktik ##able stellung ##nahme ab ##gegeben der bericht stellt einen
•	(0.91)	•	0.77	großen schrift in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs
				##timmung nach ##druck ##lich unterstutzt werden . [SEP]
forend	• Negativ	Neutral	Positive	
True	Predicted	Attribution	Attribution	fromScratch-BERT: Word Importance, checkpoint 500
Label	Label	Label	Score	[CLS] ich baha vare ##tandnis dafur, dass ain braita ##r kan ##sans baim thama umwalt ##baft ##ung nicht immar moglich ist, garada dar ##um
				mochte ich aber heute auch ganz besonders für die anna ##hme des bericht ##s des aus ##schuss ##es für recht und binnen ##markt wer ##ben
0	0	0	0.21	denn der rechts ##aus ##schuss hat eine klare "ma ##B ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben "der bericht stellt einen
	(0.96)			großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs
				##timmung nach ##druck ##lich unterstutzt werden . [SEP]
Legend	: 🔲 Negative	e 🗆 Neutral 🗖	Positive	
True	Predicted	Attribution	Attribution	fromScratch-BERT: Word Importance, checkpoint 600
Label	Laber	Laber	Score	[CLS] ich habe vers ##tandnis dafur_dass ein breite ##r kon ##sens beim thema umwelt ##haft ##ung nicht immer moglich ist_gerade dar ##um
				mochte ich aber heute auch ganz besonders für die anna ##hme des bericht ##s des aus ##schuss ##es für recht und binnen ##markt wer ##ben .
0	0	0	0.61	denn der rechts ##aus ##schuss hat eine klare, ma ##ß ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben. der bericht stellt einen
	(0.99)			großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ###m morgen bei der abs
				##timmung nach ##druck ##flich unterstutzt werden . [SEP]
egend	• Negative	Neutral	Positive	
True	Predicted	Attribution	Attribution	fromScratch-BERT: Word Importance, checkpoint 700
Label	Label	Label	Score	FOT STich halo your ##andnis dofur dage sin hesits ##e lan ##ang haim thang youry) ##haft ##your nicht journe modish ist gave do dag ##you
				[CLS] ich nabe vers ##tandnis darur, dass ein breite ##r kon ##sens beim thema umweit ##nait ##ung nicht immer möglich ist, geräde dar ##um mechta ich abar baute nuch ganz besonders für die onno ##thema der bericht ##s des ous ##cohuss ##ac für recht und binnen ##markt war ##ban
0	0	0	0.54	denn der rechts ##aus ##schuss hat eine klare ma ##R ##volle und nra ##ktik ##able stellung ##nahme ab ##gegeben der bericht stellt einen
•	(0.99)	•	0124	großen schrift in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs
				##timmung nach ##druck ##flich unterstutzt werden . [SEP]
egend	• 🗖 Negative	e 🗆 Neutral 🗖	Positive	
True	Predicted	Attribution	Attribution	fromScratch-RERT: Word Importance, checkpoint 800
Label	Label	Label	Score	
				[CLS] ich habe vers ##tandnis darur, dass ein orene ##f kon ##sens beim thema univert ##nan ##ung ment inimer möglich ist, geraue dar ##un meghte ich abar batta auch gang begendere für die anne ##hme den baiebt ##e der aus ##sehure ##tes für secht und hinnen ##meght use ##tea
0	0	0	0.43	denn der rechts ##aus ##schuss hat eine klare ma ##R ##volle und nra ##ktik ##able stellung ##nahme ab ##gegehen der bericht stellt einen
•	(0.99)		0110	großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs
				##timmung nach ##druck ##lich unterstutzt werden . [SEP]
egend	• Negative	- 🗆 Neutral 🗖	Positive	
True	Predicted	Attribution	Attribution	fromScratch-BERT: Word Importance, checkpoint 900
Label	Label	Label	Score	[C] Si joh haha vare ##tondnje dofue dage aja hrajta ##e lon ##anna kaim thoma umush ##h-0.##e====1.1.##
				[ULS] EE have vers ##tanonis datur, dass ein oreite ## kon ##sens beim thema umwelt ##hant ##ung nieht immer möglich ist, gerade dar ##um
0	0	0	0.86	denn der rechts ##aus ##schuss hat eine klare ma ##6 ##volle und pra ##ktik ##able stellung ##nahme ab ##aegeben der bericht stellt einen
-	(0.98)	,		großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m moreen bei der abs
				##unnung nach ##druck ##nch unterstutzt werden. SEP

Figure 4: Layer Integrated Gradient saliency maps of input tokens contributing to the ground truth translationese label (here: original). BERT trained from scratch for translationese classification. Changes in attribution over the training checkpoints.

Legend: True Label	■ Negative Predicted Label	□ Neutral ■ Attribution Label	Positive Attribution Score	BERT-r2c-full-ft: Word Importance, checkpoint 100
0	1 (0.72)	0	-0.28	[LLS] ich nabe vers ##tananis datur, dass ein breite ##r kon ##sens beim thema umweit ##natt ##ung nicht immer möglich ist, gerade dat ##um mochte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ##markt wer ##ben , denn der rechts ##aus ##schuss hat eine klare "ma ##B ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben .der bericht stellt einen großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs ##timmung nach ##druck ##lich unterstutzt werden . [SEP]
Legend: True	Negative Predicted	C Neutral ■ Attribution	Positive Attribution	DEDT via full for Word Importance advantaging 100
Label 0	Label 0 (0.92)	Label 0	Score -0.26	[CLS] ich habe vers ##tandnis dafur, dass ein breite ##r kon ##sens beim thema unwelt ##haft ##um micht immer moglich ist . gerade dar ##um mochte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ##markt wer ##ben , denn der rechts ##aus ##schuss hat eine klare , ma ##B ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben . der bericht stellt einen großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs ##timmung nach ##druck ##lich unterstutzt werden . [SEP]
Legend: True Label	■ Negative Predicted Label	C ■ Neutral ■ Attribution Label	Positive Attribution Score	BERT-r2c-full-ft: Word Importance, checkpoint 300
0	0 (0.99)	0	-0.02	[CLS] ich habe vers ##tandnis dafur, dass ein breite ##r kon ##sens beim thema umwelt ##haft ##ung nicht immer moglich ist. gerade dar ##um mochte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ##markt wer ##ben , denn der rechts ##aus ##schuss hat eine klare , ma ##B ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben . der bericht stellt einen großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs ##timmung nach ##druck ##lich unterstutzt werden . [SEP]
Legend True	■ Negative Predicted	e □ Neutral ■ Attribution	Positive Attribution	RFRT-r2c-full-ft: Word Importance, checkpoint 400
Label	Label	Label	Score	[CLS] ich habe vers ##tandnis dafur, dass ein breite ##r kon ##sens beim thema umwelt ##haft ##ung nicht immer moglich ist .gerade dar ##um
0	0 (0.93)	0	0.01	mochte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ##markt wer ##ben, denn der rechts ##aus ##schuss hat eine klare, ma ##β ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben. der bericht stellt einen großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs ##timmung nach ##druck ##lich unterstutzt werden. [SEP]
Legend True	■ Negative Predicted	Neutral	Positive Attribution	REPT at full for Ward Importance a bashroot 200
Label 0	Label 0 (0.98)	Label 0	Score -0.07	[CLS] ich habe vers ##tandnis dafur, dass ein breite ##r kon ##sens beim thema umwelt ##haft ##ung nicht immer moglich ist "gerade dar ##um mochte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ##markt wer ##ben , denn der rechts ##aus ##schuss hat eine klare , ma ##B ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben . der bericht stellt einen großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs ##timmung nach ##druck ##lich unterstutzt werden . [SEP]
Legend True	■ Negative Predicted	• 🗆 Neutral 🔲	Positive Attribution	
Label 0	Label 0 (0.98)	Label 0	Score -0.57	ICLS] ich habe vers ##tandnis dafur, dass ein breite ##r kon ##sens beim thema umwelt ##haft ##ung nicht immer moglich ist .gerade dar ##um mochte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ##markt wer ##ben , denn der rechts ##aus ##schuss hat eine klare , ma ##ß ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben . der bericht stellt einen großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch/im pl ##enu ##m morgen bei der abs ##timmung nach ##druck ##lich unterstutzt werden . [SEP]
Legend True	■ Negative	Neutral	Positive Attribution	
Label	Label	Label	Score	BERT-r2c-full-ft: Word Importance, checkpoint 700
0	0 (0.98)	0	-0.70	The nucle real manufacture and the set of th
Legend True	Negative Predicted	e □ Neutral ■ Attribution	Positive Attribution	DEPT via full fr: Word Importance absolute 200
Label 0	Label 0 (0.97)	Label 0	Score -0.42	[CLS] ich habe vers ##tandnis dafur, dass ein breite ##r kon ##sens beim thema umwelt ##haft ##ung nicht immer moglich ist .gerade dar ##um mochte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ##markt wer ##ben , denn der rechts ##aus ##schuss hat eine klare , ma ##B ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben . der bericht stellt einen großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch/im pl ##enu ##m morgen bei der abs ##timmung nach ##druck ##lich unterstutzt werden . [SEP]
Legend True	■ Negative Predicted	Neutral Attribution	Positive Attribution	REPT-r2.cfull-fr: Word Importance elsestanist 000
Label 0	Label 0 (0.97)	Label 0	Score -0.35	[CLS] ich habe vers ##tandnis dafur, dass ein breite ##r kon ##sens beim thema umwelt ##haft ##ung nicht immer moglich ist . gerade dar ##um mochte ich aber heute auch ganz besonders fur die anna ##hme des bericht ##s des aus ##schuss ##es fur recht und binnen ##markt wer ##ben , denn der rechts ##aus ##schuss hat eine klare , ma ##B ##volle und pra ##ktik ##able stellung ##nahme ab ##gegeben . der bericht stellt einen großen schritt in richtung eines verb ##esse ##rten umwelt ##schutz ##es dar und sollte deshalb auch im pl ##enu ##m morgen bei der abs ##timmung nach ##druck ##lich unterstutzt werden . [SEP]

Figure 5: Layer Integrated Gradient saliency maps of input tokens contributing to the ground truth translationese label (here: original). BERT pretrained for handcrafted feature prediction, and fine-tuned for translationese classification. Changes in attribution over the training checkpoints.

Legend: True Label	Negative Predicted Label	□ Neutral ■ I Attribution Label	Positive Attribution Score	fromScratch-BERT: Word Importance, checkpoint 100
1	1 (0.54)	1	-1.06	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##au ##men "wir mussen in zwei richtung ##en vor ##gehen "kurz – und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##sso ##ue ##rani ##tat lege ##n "zwei ent ##scheiden ##de aspekte "um den heute existieren ##den hunger zu mild ##erm . langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP]
Legend: True	Negative Predicted	□ Neutral ■ I Attribution	Positive Attribution	fromScratch-BERT: Word Importance, checkpoint 200
Label	Label 1 (0.55)	Label	Score 0.60	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##au ##men "wir mussen in zwei richtung ##en vor ##gehen "kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##si ##cher ##heit und die er #mah ##rung ##so ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte "um den heute existieren ##den hunger zu mild ##ern "langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden[SEP]
Legend: True Label	Negative Predicted Label	□ Neutral ■ I Attribution Label	Positive Attribution Score	fromScratch-BERT: Word Importance, checkpoint 300
1	0 (0.57)	1	1.19	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##au ##men "wir mussen in zwei richtung ##en vor ##gehen "kurz – und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte "um den heute existieren ##den hunger zu mild ##erm "langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden "[SEP]
Legend: True	Negative Predicted	Neutral	Positive Attribution	fromScratch-BERT: Word Importance, checkpoint 400
Label	Label 0 (0.56)	Label 1	Score 0.81	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##men [wir mussen in zwei richtung ##en vor ##gehen]kurz =und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte , um den heute existieren ##den hunger zu mild ##ern]anger ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden [[SEP]
Legend: True	Negative Predicted	□ Neutral ■ I Attribution	Positive Attribution	from Constab. DEDT: Word Importance advaluation 500
Label	Label 1 (0.67)	Label	Score 3.56	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##au ##men "wir mussen in zwei richtung ##en vor ##gehen "kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n "zwei ent ##scheiden ##de aspekte "um den heute existieren ##den hunger zu mild ##em "langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden "[SEP]
Legend: True Label	Negative Predicted Label	□ Neutral ■ I Attribution Label	Positive Attribution Score	fromScratch-BERT: Word Importance, checkpoint 600
1	1 (0.71)	1	2.44	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##au ##men "wir mussen in zwei richtung ##en vor ##gehen "kurz – und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##eher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n, zwei ent ##scheiden ##de aspekte "um den heute existieren ##den hunger zu mild ##ern "langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden[SEP]
Legend: True	Negative Predicted	Neutral Attribution	Positive Attribution	fromScratch-BERT: Word Importance, checkpoint 700
Label	Label 1 (0.75)	Label 1	3.02	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##au ##men "wir mussen in zwei richtung ##en vor ##gehen "kurz -und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n "zwei ent ##scheiden ##de aspekte "um den heute existieren ##den hunger zu mild ##ern "langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP]
Legend: True	■ Negative Predicted	Neutral	Positive Attribution	fromScratch-BERT: Word Importance, checkpoint 800
Label	Ladei 1 (0.82)	Ladel	3.34	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung vor ##rang ein ##zur ##au ##men , wir mussen in zwei richtung ##en vor ##gehen , kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte , um den heute existieren ##den hunger zu mild ##ern , langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP]
Legend: True	Negative Predicted	□ Neutral ■ I Attribution	Positive Attribution	fromScratch-BERT: Word Importance, checkpoint 900
Label 1	Label 1 (0.81)	Label	Score 3.26	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##au ##men "wir mussen in zwei richtung ##en vor ##gehen "kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n "zwei ent ##scheiden ##de aspekte "um den heute existieren ##den hunger zu mild ##em "langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um für die weniger entwickelten

Figure 6: Layer Integrated Gradient saliency maps of input tokens contributing to the ground truth translationese label (here: translation). BERT trained from scratch for translationese classification. Changes in attribution over the training checkpoints.

Legend True Label	: ■ Negative Predicted Label	□ Neutral ■ I Attribution Label	Positive Attribution Score	BERT-r2c-full-ft: Word Importance, checkpoint 100
1	1 (0.76)	1	1.11	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet " der menschlichen entwicklung vor ##rang ein ##zur ##au ##men "wir mussen in zwei richtung ##en vor ##gehen "kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n " zwei ent ##scheiden ##de aspekte " um den heute existieren ##den hunger zu mild ##ern "langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden " [SEP]
Legend True Label	: Negative Predicted Label	□ Neutral ■ I Attribution Label	Positive Attribution Score	BERT-r2c-full-ft: Word Importance, checkpoint 200
1	0 (0.82)	1	0.98	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet "der menschlichen entwicklung vor ##rang ein ##zur ##au ##men "wir mussen in zwei richtung ##en vor ##gehen "kurz – und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte "um den heute existieren ##den hunger zu mild ##ern "langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP]
Legend True Label	: ■ Negative Predicted Label	□ Neutral ■ I Attribution Label	Positive Attribution Score	BERT-r2c-full-ft: Word Importance, checkpoint 300
1	1 (0.93)	1	2.64	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet, der menschlichen entwicklung vor ##rang ein ##zur ##au ##men , wir mussen in zwei richtung ##en vor ##gehen , kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte , um den heute existieren ##den hunger zu mild ##ern , langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP]
Legend True	: Negative Predicted	□ Neutral ■ I Attribution	Positive Attribution	BERT-r2c-full-ft: Word Importance, checkpoint 400
Label	Label	Label	Score	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung
1	1 (0.91)	1	2.69	vor ##rang ein ##zur ##au ##men "wir mussen in zwei richtung ##en vor ##gehen "kurz – und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte , um den heute existieren ##den hunger zu mild ##ern , langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden "[SEP]
Legend True	: Negative Predicted	Neutral I Attribution	Positive Attribution	DEDT eta full fri Ward Importanza abaskualat 500
Label	Label 1 (0.70)	Label 1	Score 2.84	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet, der menschlichen entwicklung vor ##rang ein ##zur ##au ##men , wir mussen in zwei richtung ##en vor ##gehen , kurz = und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##sis i##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani i#tat lege ##n , zwei ent ##scheiden ##de aspekte , um den heute existieren ##den hunger zu mild ##ern , langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP]
Legend True Label	: ■ Negative Predicted Label	□ Neutral ■ I Attribution Label	Positive Attribution Score	BERT-r2c-full-ft: Word Importance, checkpoint 600
1	1 (0.75)	1	3.27	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet, der menschlichen entwicklung vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##en vor ##gehen . kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte , um den heute existieren ##den hunger zu mild ##em . langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP]
Legend True	: Negative Predicted	□ Neutral ■ I Attribution	Positive Attribution	
Label	Label	Label	Score	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet , der menschlichen entwicklung
1	1 (0.71)	1	3.61	vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##en vor ##gehen . kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte , um den heute existieren ##den hunger zu mild ##ern . langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP]
Legend True	: Negative Predicted	□ Neutral ■ I Attribution	Positive Attribution	BERT-r2c-full-ft: Word Importance, checkpoint 800
Label	Label 1 (0.71)	Label	Score 3.52	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet, der menschlichen entwicklung vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##en vor ##gehen . kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte , um den heute existieren ##den hunger zu mild ##ern . langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP]
Legend True Label	■ Negative Predicted Label	□ Neutral ■ I Attribution Label	Positive Attribution Score	BERT-r2c-full-ft: Word Importance, checkpoint 900
1	1 (0.81)	1	3.59	[CLS] das problem an der wur ##zel zu pack ##en und die gegenwart ##ige krise zu uber ##wind ##en bedeutet, der menschlichen entwicklung vor ##rang ein ##zur ##au ##men . wir mussen in zwei richtung ##en vor ##gehen . kurz - und mittel ##fr ##isti ##g mussen wir den schwerpunkt auf die er ##nah ##rung ##ssi ##cher ##heit und die er ##nah ##rung ##sso ##uve ##rani ##tat lege ##n , zwei ent ##scheiden ##de aspekte , um den heute existieren ##den hunger zu mild ##em . langer ##fr ##isti ##g muss das nach ##halt ##ige wachst ##um fur die weniger entwickelten lander gefordert werden . [SEP]

Figure 7: Layer Integrated Gradient saliency maps of input tokens contributing to the ground truth translationese label (here: translation). BERT pretrained for handcrafted feature prediction, and fine-tuned for translationese classification. Changes in attribution over the training checkpoints.