How Language-Dependent is Emotion Detection? Evidence from Multilingual BERT

Luna De Bruyne[*], **Pranaydeep Singh**[*], **Orphée De Clercq**,
Els Lefever, **Véronique Hoste**
LT^3, Language and Translation Technology Team, Ghent University
`{firstname.lastname}@ugent.be`

Abstract

As emotion analysis in text has gained a lot of attention in the field of natural language processing, differences in emotion expression across languages could have consequences for how emotion detection models work. We evaluate the language-dependence of an mBERT-based emotion detection model by comparing language identification performance before and after fine-tuning on emotion detection, and performing (adjusted) zero-shot experiments to assess whether emotion detection models rely on language-specific information. When dealing with typologically dissimilar languages, we found evidence for the language-dependence of emotion detection.

1 Introduction

As language finds itself at the crossroads of cognition and culture, it has been a thoroughly investigated subject in the context of emotion research and has given rise to questions as how emotion expression varies across languages and whether language has an impact on emotion conceptualisation and perception.

Indeed, many studies have reported on the cultural relativity of emotion, often underscoring the diversity in emotion lexicons across languages: not only is there a big variability in which emotional states are included in the lexicon of a language with a designated emotion term (e.g., a word for *sadness* seems to be missing in Tahiti (Levy, 1984)), but there are also many differences in the connotations and meanings of emotion terms across languages (Mesquita et al., 1997; Wierzbicka, 1999).

Instead of focusing on emotion conceptualisation and experience, one could also ask whether emotions are *expressed* differently across languages. Again, this can be reflected in differences in emotion vocabulary, but also in language-specific phraseology. In Russian, for example, the verbalisation of emotion is very much focused on the human body, and the numerous diminutive suffixes exhibit different emotional nuances (Wierzbicka, 1999). Noteworthy is also the distinction between individualistic and collectivist cultures, where the latter are associated with more reticence to express emotions, while the former exhibit more open emotion expression (Semin et al., 2002).

As emotion analysis in text has gained a lot of attention in artificial intelligence and the field of natural language processing (NLP) as well (Calvo and Mac Kim, 2013; Mohammad, 2016), language-dependent conceptualisation and expression could have consequences for how emotion detection models work. Analogously to humans who might need knowledge about the linguistic code (e.g., to know whether irony is often used in a specific language or to understand language-specific phraseology) to correctly judge the emotional value of someone's utterance, machine learning models might need this knowledge as well in order to accurately predict emotions from text. Therefore, we investigate the language-dependence of the task of emotion detection. In other words, we want to know whether knowledge about the language identity is needed to make accurate emotion predictions.

For this analysis, we will look at languages from different language families and branches (e.g., Germanic, Italic and Indo-Iranian in the Indo-European language family or Chinese from the Sino-Tibetan language family) in order to include languages with different structural features. Although language families are not the same as the classes defined in the field of linguistic typology (i.e., the analysis, comparison, and classification of languages according to their common structural features and forms), languages within one language family are generally more typologically similar than languages from different families.

As transformer models are currently state of the

[*] These authors contributed equally to this work.

art in many NLP tasks, we investigate the language-dependence of multilingual BERT (mBERT), the transformer model introduced by Devlin et al. (2019) which was trained on 104 languages. We foresee two kinds of experiments. First, we investigate how much language-specific information is preserved in the BERT representations by comparing performance on the task of language identification both before and after fine-tuning on emotion detection. Second, zero-shot transfer learning (training on a source language and testing directly on the target language English) is compared with training on machine-translated data, i.e., data that was originally in English but automatically translated to the source language ('semi-zero-shot transfer learning'). These models thus learn from the same source language, but in the semi-zero-shot set-up language-specific information from the target language (like idioms, phraseology or cultural codes) might still be preserved, thus aiding performance during test time on the target language.

In Section 2, we describe the literature on cross-lingual emotion research (Section 2.1) and discuss related work dealing with language dependency in NLP (Section 2.2). In Section 3 we explain our method by describing the data and resources (Section 3.1) and by zooming in on the experimental set-up (Section 3.2). The results are reported in Section 4 and further discussed in Section 5, followed by a conclusion in Section 6.

2 Related work

2.1 Emotions across languages

While many psychological models assume that emotions are distinct from linguistic processing, growing psychological research suggests that language plays an important role in both emotion experience and perception. Especially in psychological constructionist theories of emotion, language is considered as doing more than merely communicating emotion. Instead, language contributes to the conceptualisation of emotion itself (Lindquist et al., 2015).

In the constructionist view, the experience of emotion takes place when sensations inside and outside the body are made meaningful in their context by use of concept knowledge. This is referred to as the theory of constructed emotion or – as it was previously called – the conceptual act theory (Barrett, 2006). Concept knowledge is the knowledge we have about different categories, acquired via semantic knowledge and personal experience (Lindquist et al., 2015). Both language and culture can thus play an important role here.

The role of language in emotion can be linked to the linguistic relativity hypothesis (Whorf, 1956). Linguistic relativity, often referred to as the Sapir-Whorf hypothesis, suggests that the way people think is influenced by the language they speak. Speakers of Russian, for example, a language which has separate words for naming light blue (*goluboy*) and dark blue (*siniy*), discriminate between various shades of blue differently than English speakers, who only have one term to denote blue (Winawer et al., 2007). Another example of linguistic relativity is the observation that Inupiaq, an Inuit language, has many words for snow, while English has only one, which suggests that speakers of these languages categorize their environment differently. In this light, it is compelling to study cross-lingual differences in emotion conceptualisation, experience and perception.

In the context of emotion conceptualisation, Mesquita et al. (1997) highlighted that lexical equivalents are mostly not expressing the same meaning across languages. This is in line with results from a colexification analysis of emotion words in 2,474 languages, in which Jackson et al. (2019) found that there is a wide variation in which emotion concepts are lexicalized together by one word form, and that colexifications vary systematically across language families. In Tai-Kadai languages, for example, *anxiety* is closely related to *fear*, while it is more related to *grief* and *regret* in Austroasiatic languages.

Also emotion perception varies across languages, which is reflected in differences in emotionality ratings (affective norms) of words (Harris et al., 2006). Of course, this could be linked to the differences in meaning in lexical equivalents across languages, but it might also be due to cultural differences in appraisal of the same event. Mesquita and Ellsworth (2001) give as example that solitude may be perceived as positive in middle-class European culture and lead to *contentment*, while in Inuit culture, being alone is typically associated with *sorrow* and for Tahitians with *fear*.

Finally, there is also variation in how emotions are expressed. Semin et al. (2002) found that individualistic cultures and collectivist cultures express emotions and emotional events using different linguistic markers and divergent levels of abstraction: in individualistic cultures, emotion terms are more prominent as self-markers and are represented by abstract language (e.g., adjectives and nouns), while in collectivist cultures, emotion terms are more prominent as relationship-markers and are represented by concrete language (e.g., interpersonal verbs). This is in line with studies on emotional reticence in East Asian cultures. Caldwell-Harris et al. (2013) compared verbal declarations of *love* in Chinese and American English, where they placed the reticence of both verbal and non-verbal emotional expression in Chinese opposite to the frequent use of 'I love you' as displaying American expressivity.

2.2 Language dependency in natural language processing

Cross-lingual and multilingual perspectives on natural language processing have received a lot of attention, especially regarding the transferability of NLP models across languages. Since the rise of deep learning, many efforts have been made to achieve cross-lingual representations of words in a joint embedding space (Ruder et al., 2019). Also state-of-the-art transformer models have been developed in multilingual variants, like multilingual BERT (mBERT) (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2020).

These multilingual models have been the subject of probing studies to investigate how well they perform in zero-shot cross-lingual model transfer (i.e., fine-tuning the model on task-specific training data from a source language and testing that resulting model on test data for that task in a different language). Pires et al. (2019) performed such probing experiments with mBERT (named entity recognition and part of speech tagging) and found that it has a robust ability to generalize cross-lingually, but that transfer works best between typologically similar languages. This could indicate that mBERT learns representations which contain both a cross-lingual and a language-specific component. Using Canonical Correlation Analysis (CCA) on the internal representations of mBERT, Singh et al. (2019) found that mBERT is not embedding different languages into one shared space, but that it partitions representations for each language (especially at deeper layers) in a way that reflects the linguistic and evolutionary relationships between languages as represented in phylogenetic trees. When looking at the representations of the last layer of mBERT, Gonen et al. (2020) could identify a language-identity subspace, which supports the hypothesis that there are identifiable language components in mBERT.

While there are many studies trying to gain insight in how language-specific information is stored in mBERT, the focus is mostly on the embeddings themselves, and not on how different tasks exploit this information. An exception is the study of Tanti et al. (2021), who investigated the effect of fine-tuning on specific tasks on the language-specific component of mBERT representations. They found that mBERT's representations become less language-specific after fine-tuning and that there is a greater loss of this information in POS-tagging, which is a morphosyntactic task, compared to natural language inference (NLI), which is a semantically oriented task.

For the task of emotion detection, the exploitation of language-specific information in word embeddings has not yet been investigated. However, language-dependence of this task and the related task of sentiment analysis has been studied in the context of emotion/sentiment preservation after translation. Mohammad et al. (2016) investigated the use of Support Vector Machines in detecting sentiment (positive/negative/neutral) in Arabic social media posts and compared performance of an Arabic sentiment classification system with an English system where the Arabic texts were translated to English. They found that the translation-based approach produced results on par with Arabic sentiment analysis when the translation was done manually, and led to a small drop in performance when the translation was done automatically. This suggests that, when using high-quality translations, sentiment analysis does not suffer from losing language-specific information. However, the authors did observe that translations often did not preserve the original sentiment and investigated this by means of an annotation task of the instances where translation had resulted in sentiment change. When the translation was done automatically, the main reason for sentiment change was bad translation, but when the translation was done manually, the annotators indicated cultural differences as the main reason for this change. An example of the latter is a sentence that referred to not seeing the crescent moon and that was annotated in English as neutral, but negative in Arabic, as the crescent moon in Islam is associated with the beginning of a month or a holiday. Another example included the phrase "I have no comment", which was annotated as neutral in English, but is used to express a negative opinion in Arabic.

A similar study was performed by Kajava et al. (2020), who investigated the preservation of the emotion categories *anger, anticipation, disgust, fear, joy, sadness, surprise* and *trust* when English data was translated to Finnish, French and Italian. They deemed the degree of preservation sufficient for using translated data in cross-lingual emotion detection systems and found that change of emotion labels was due to incomplete or ambiguous translation and to the difficulty of the emotion annotation task itself (which even causes confusion between the annotations of annotators within one language), rather than to linguistic differences in the encoding of emotion.

3 Method

3.1 Resources & data

We assess the language dependency of emotion detection using multilingual BERT (mBERT), which was released together with the original English BERT model (Devlin et al., 2019). Both the English and multilingual BERT are 12-layer transformers, but while the original BERT is trained on English data only, mBERT is trained on the Wikipedia pages of 104 languages and thus has a shared word piece vocabulary. There is no explicit marker denoting the input language, nor does it use an explicit mechanism to encourage translation-equivalent pairs to have similar representations.

Emotion dataset For our emotion detection dataset, we start from the Universal Joy dataset (Lamprinidis et al., 2021). The original dataset consists of 530k Facebook posts in 18 languages, which were collected based on the 'feelings tags' that users added to their message. These self-labeled tags were then mapped to one of the 5 emotion categories *anger*, *anticipation*, *fear*, *joy* and *sadness*. For our experiments, we included all languages from the 'Small' version of this dataset (2,947 instances per language), namely Chinese, English, Portuguese, Spanish and Tagalog, and complemented this with the Dutch (as it is typologically very similar to English) and Hindi (to have an additional more typologically distinct language) data from the 'Low Resource' subset (2,201 instances for Dutch and 1,823 for Hindi).

We made sure the sizes of the datasets and distributions of the emotion labels were equal across all seven languages, which will be important for the zero-shot experiments (see Section 3.2). We therefore identified the language with the lowest number of instances for each label, and randomly sampled the same number of instances with that label for the other languages. This resulted in 10,437 instances in total or 1,491 instances per language, of which 150 for *anger*, 231 for *anticipation*, 8 for *fear*, 830 for *joy* and 272 for *sadness*. We call this set `UJ Equal`. The original Universal Joy dataset contains some special tokens like `[URL]`, `[PHOTO]`, `[LOCATION]` or `[PERSON]`. We removed all of these except `[PERSON]`, as they are not part of the grammatical sentence.

We also provide a dataset with machine translations, based on the English part of `UJ Equal`. Using the Google Translate API with the Python package `googletrans`[1], we translated the English subset in `UJ Equal` to Chinese, Dutch, Hindi, Portuguese, Spanish and Tagalog and call this dataset `UJ MT`.

We further have a separate test set of English instances consisting of 981 sentences, as provided in the original Universal Joy dataset, which we call `UJ English Test`.

Language Identification dataset 6,000 instances for each of the seven languages (Chinese, Dutch, English, Hindi, Portuguese, Spanish and Tagalog) were taken from the OSCAR corpus (Ortiz Suárez et al., 2020), which is a multilingual corpus obtained by language classification and filtering of the Common Crawl corpus[2]. These instances were randomized and the language code was added as label.

3.2 Experimental setup

Preservation of language-specific information First, we investigate to what degree language-specific information is preserved after fine-tuning mBERT on the task of emotion detection. We use the pre-trained mBERT model with a single-layer softmax classifier on top. In phase 1, the pre-trained model is used without fine-tuning to execute the language identification task (7-class classification on the Language Identification dataset). In phase 2, mBERT is fine-tuned in 5 epochs on the

[1] https://pypi.org/project/googletrans/

[2] https://commoncrawl.org/

emotion detection task (with the 10,437 instances from UJ Equal) using categorical cross-entropy loss. The resulting model is then used for the encoding and classification of the language identification task. The language identification performance of both phases is then compared. Moreover, we visualize the outputs from different layers in the BERT model, and at different stages in the fine-tuning process using t-SNE to decipher the effect of fine-tuning for emotion on the language-specific representations.

Zero-shot and semi-zero-shot experiments The next set of emotion detection experiments also consists of two phases. In phase 1, more traditional zero-shot experiments are performed, where we either train on the source languages Chinese, Dutch, Hindi, Portuguese, Spanish or Tagalog (1,491 instances from UJ Equal), and test on the separate English set of 981 sentences (UJ English Test). In phase 2, we train on the same source languages, but instead of relying on authentic, original data we rely on the UJ MT data. This data was thus originally in English, but machine-translated to either Chinese, Dutch, Hindi, Portuguese, Spanish or Tagalog. We call this semi-zero-shot experiments.

The idea behind this is that the machine-translated data could be closer to the target language regarding language-specific information, and that the version with machine-translated data will thus perform better in tasks where language-specific information is important. Note that it is crucial that all train sets have the same label distribution, to avoid that the (dis)similarity with the label distribution of the test set explains the performance of the models.

Again, we use pre-trained mBERT with a single-layer softmax classifier and cross-entropy loss as loss function. We compare the (semi-)zero-shot models against a within-language baseline, trained on the English part of the UJ Equal dataset.

4 Results

4.1 Preservation of language-specific information

Effect on language identification performance

The language identification performance before and after fine-tuning on emotion detection is shown in Table 1. When using the pre-trained mBERT model without further fine-tuning, the model achieves a macro-averaged F1-score of 99.92%. This means that mBERT reaches an almost perfect performance in differentiating between languages, which is in line with previous findings that mBERT partitions representations per language (Singh et al., 2019) or that it at least exhibits a language-identity subspace (Gonen et al., 2020).

Task	Macro F1
before fine-tuning (frozen LM)	0.9992
after fine-tuning on emotion detection	0.9161

Table 1: Language identification performance before and after fine-tuning on emotion detection.

When fine-tuning mBERT on emotion detection and applying the resulting model to perform language identification, the model's performance drops to 91.61%. As also observed by Tanti et al. (2021), the mBERT representations become less language-specific after fine-tuning on a specific task. Intuitively, tasks that require less language-specific knowledge, would lose more language-specific information than tasks that heavily rely on language-specific knowledge, resulting in a larger drop of language identification performance. As the drop in performance after fine-tuning on emotion detection (7.47%) is relatively small (especially compared to the drops reported by Tanti et al. (2021), which was 10.6% after fine-tuning on NLI and even 78% for POS-tagging), one could deduce that emotion detection does rely rather heavily on language-specific knowledge.

T-SNE plots

To visualise the effect of fine-tuning for emotion detection on the mBERT representations, the hidden states of the first (Layer 1), middle (Layer 6) and last (Layer 12) layer of the model are plotted in Figure 1 using t-SNE projections before fine-tuning (Epoch 0), and after Epoch 2 and 4.

We see that, regardless of how far the fine-tuning process has progressed, the languages are already clearly distinct in the first layer of the model. In the last layers, the language clusters begin to slowly merge while the model is being fine-tuned.

After epoch 2, most languages have already merged, but Chinese, Hindi and Tagalog (the non-European languages) are still represented in separate clusters. However, after epoch 4, Hindi and Tagalog have entered the European cloud, while Chinese stays more or less isolated.

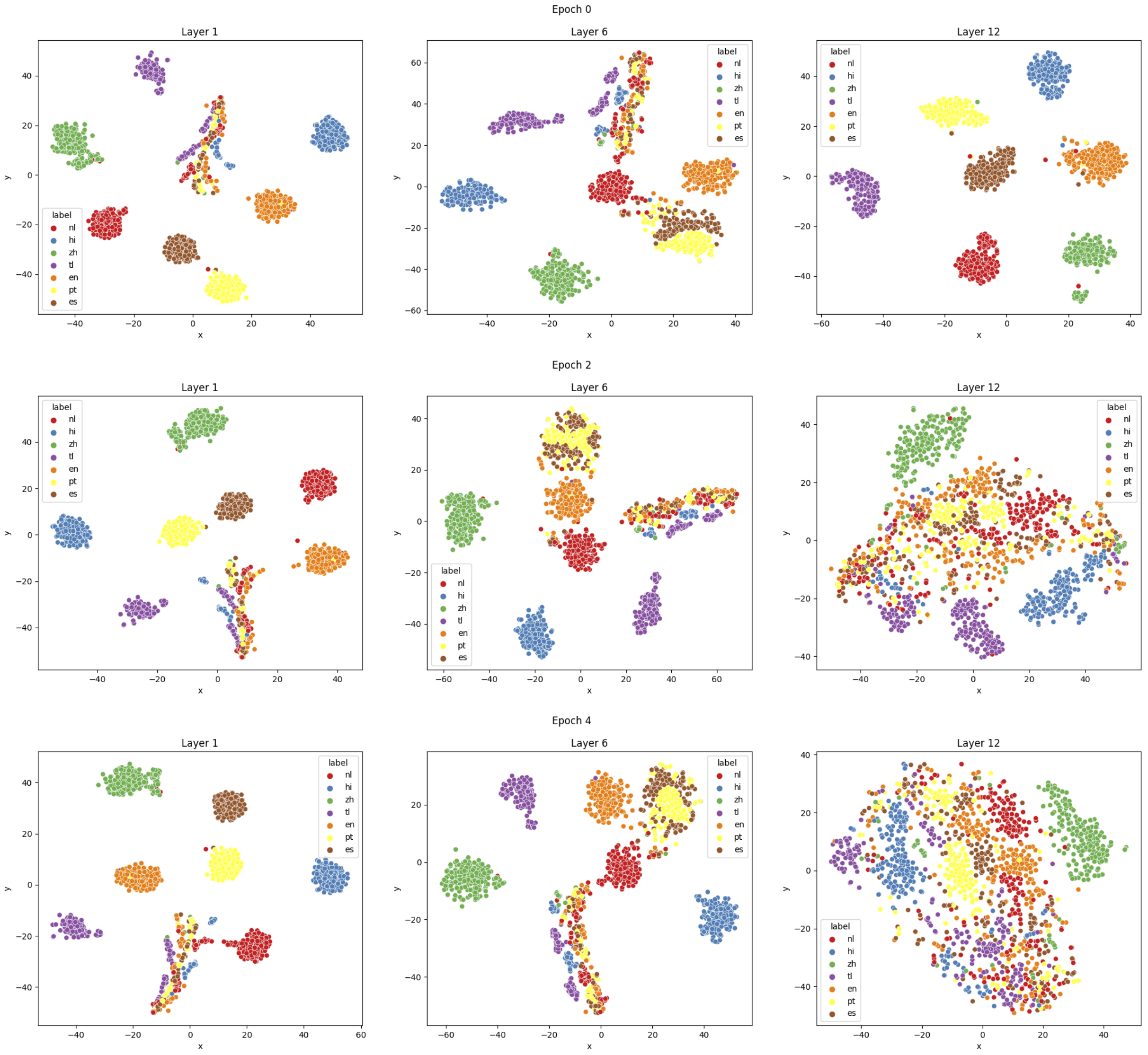

Figure 1: Visualisation of mBERT embeddings and effect on language separation when fine-tuning on emotion detection. *Language codes*: nl = Dutch, hi = Hindi, zh = Chinese, tl = Tagalog, en = English, pt = Portuguese, es = Spanish.

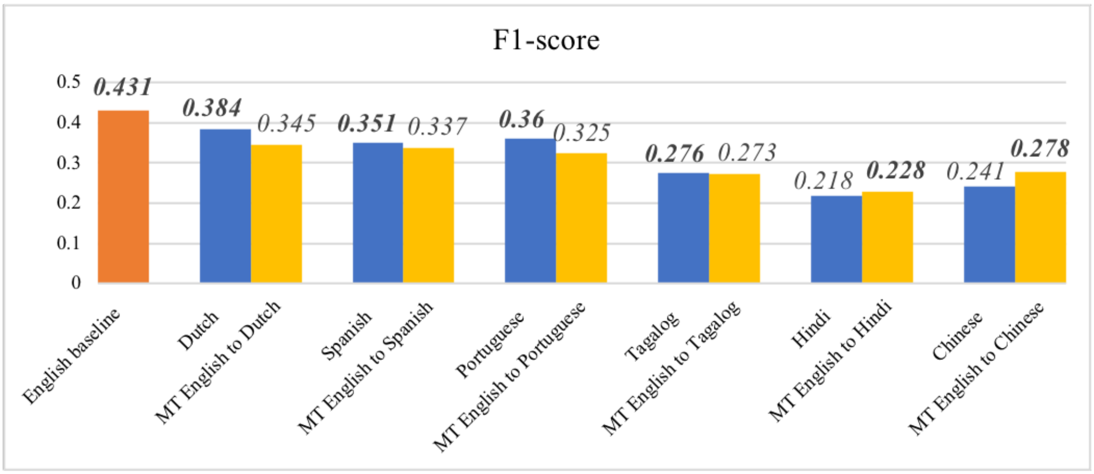

Figure 2: F1-scores for zero-shot (blue) and semi-zero shot (yellow) emotion classification on the English test set.

Interestingly, instead of a complete mix-up of languages, each language is still quite distinguishable after fine-tuning, even though they are being placed much closer to each other than initially. Especially Chinese and Hindi, but also Portuguese are easily distinguishable.

4.2 Zero-shot and semi-zero-shot experiments

In this section, we report zero-shot and semi-zero-shot results for emotion detection on the `UJ English test` set from Universal Joy. The baseline macro-average F1-score (trained on the English part of `UJ Equal`) is 43.1%.

As shown in Figure 2, all zero-shot experiments achieve a lower performance than this baseline, with Hindi as lowest performing source language (21.8% F1), followed by Chinese (24.1% F1) and Tagalog (27.6% F1). Unsurprisingly, Dutch is the best-performing source language (38.4% F1), followed by Portuguese (36.0% F1) and Spanish (35.1% F1). The performance of these source languages more or less corresponds to their typological similarity, with the European languages performing best as source language when English is the target language. Only Hindi, which also belongs to the family of Indo-European languages, performs worse than expected (even worse than Chinese and Tagalog, which belong to the Sino-Tibetan and Austronesian language families, respectively). This might be due to the difference in script (Devanagari for Hindi versus Latin for the other Indo-European languages).

Our idea was that using machine-translated instances (English to source language) as training data instead of real instances in the source language, would give an indication of the system's reliance on language-specific information, as some of this information might still be preserved in a (machine) translation. Before the translation step, all training instances in these so-called semi-zero-shot experiments are the same, namely the English part of `UJ equal`. We expected a drop in the semi-zero-shot results compared to the baseline results (because some information will be lost anyway due to (imperfect) translation), but if the drop from baseline to semi-zero-shot would be smaller compared to the drop from baseline to normal zero-shot, this might indicate that the model relies more on language-specific information (note that the size of the fine-tuning set is equal in the zero-shot experiments and semi-zero-shot experiments). These results are indicated by the yellow bars in Figure 2.

Interestingly, we see that for the European languages, normal zero-shot is better than semi-zero-shot (with normal zero-shot outperforming semi-zero-shot with around 4 to 6% F1), while for Chinese and Hindi semi-zero-shot is better. The results for Tagalog are less outspoken, as the F1-score for zero-shot (27.6%) and semi-zero-shot (27.3%) are on par.

If it is the case that language-specific information is really encoded in the machine-translated instances, then these results could indicate that an emotion detection model does rely on such information. The language-specific information might be similar for English and the other European languages used in this study, making that there is no benefit in using a model that encodes this information for English (i.e., the semi-zero-shot model). However, for less similar languages, these results do suggest that there is a benefit and that emotion detection is language-dependent.

5 Discussion

Although we found some potential evidence for the language-dependence of emotion detection, several points need to be taken into account. First of all, the datasets used in this study are small (especially for the category *fear*), and the overall quality of the data is low. It seems that some messages are incomplete and that some (parts of) instances appear multiple times in the dataset.[3] Furthermore, some instances contain code-switching between different languages. Another drawback is that we only tested on English. We made this choice because we could not obtain test sets for all languages (for Hindi and Dutch, all data was already used for training).

We claim that we found evidence for the language-dependence of emotion detection, where typologically dissimilar languages suffer more from cross-lingual zero-shot learning. This evidence is partly based on the observation that semi-zero-shot experiments (in which language-specific

[3] Example from the Dutch subset of Universal Joy:*"valiumpilletje gekregen om rustig te worden, haar lichaam moet de rest doen, maar de eerste uren heeft ze zich er ernstig tegen verzet maar ligt nu gelukkig heerlijk te slapen.* ***Hopelijk voor ons allen een goede [PERSON] .*** *"; "tegen verzet maar ligt nu gelukkig heerlijk te slapen* ***. Hopelijk voor ons allen een goede*** *nachtrust."; "heerlijk te slapen* ***. Hopelijk voor ons allen een goede [PERSON] .*** *"; "slapen.* ***Hopelijk voor ons allen een goede [PERSON] .*** *"; "* ***. Hopelijk voor ons allen een goede [PERSON] .*** *"* are separate instances in the dataset.

information is assumed to be preserved to a certain extent) outperforms zero-shot learning for Hindi and Chinese, while it does not for the European languages (as language-specific information might be similar for these languages and English, and there therefore is no benefit in using a model that encodes this information). However, it could be that this language-specific information is not related to phraseology or differences in emotion-topic relations (see Section 1 and 2.1), but to differences in topic distribution in general. It might be the case that the topics in Chinese and Hindi are very different from the topics in the English dataset, while the European languages contain similar topic distributions as English.

The semi-zero-shot experiments are based on the idea that some language-specific information is preserved after machine translation. Although we cannot be absolutely certain of this, the fact that the semi-zero-shot experiments outperformed normal zero-shot for some languages, suggests that there is some helpful information in these translations. One could argue that the performance of the semi-zero-shot models correlates negatively with the quality of the translations: machine translation might be bad for less similar languages, resulting in a better emotion classification performance in the semi-zero-shot case, because some words have not been translated. However, we could not find evidence for this. When applying a token-level language identifier on the translated texts[4], we found that the percentage of tokens that was classified as English instead of Chinese, Tagalog and Hindi is respectively 6%, 3% and 0.3%. That there are almost no untranslated words in the Hindi set while the semi-zero-shot does perform better, thus contradicts that the semi-zero-shot performance is explained by the number of untranslated words.

In future work, we envisage to use a different approach for investigating the language-dependence of emotion detection instead of relying on semi-zero-shot experiments. As both this study and previous research has shown that mBERT partitions its representations per language (Singh et al., 2019; Gonen et al., 2020), it would be compelling to see whether we can achieve language-neutral representations and which effect that has on the emotion detection performance. We hypothesise that when the representations no longer exhibit language-specific information, it would hamper emotion detection. However, in such a set-up, we will need to compare emotion detection to a reference task and discuss the language dependency of those tasks in relation to each other. This because the process of making language-neutral representations will involve reducing the transformer's parameters and that will probably lead to a performance drop anyway.

6 Conclusion

In this paper, we assessed the language-dependence of an mBERT-based emotion detection model. We first investigated the effect of fine-tuning on emotion on the preservation of language-specific information in mBERT, by comparing language identification performance of the languages Chinese, Dutch, English, Hindi, Portuguese, Spanish and Tagalog before and after fine-tuning on emotion detection and visualising the model's hidden states in t-SNE plots. As expected, language-specific information is lost after fine-tuning, but only to a small extent. Especially the representations of typologically dissimilar languages remain more or less isolated, while similar languages get clustered together.

In a next set of experiments, we compared zero-shot learning with what we called 'semi-zero-shot learning'. In the zero-shot experiments, we trained a model on either Chinese, Dutch, Hindi, Portuguese, Spanish or Tagalog and tested it on English data. In semi-zero-shot, originally English data was translated to those languages, assuming that some language-specific information is preserved in these translations. We found that for the European languages, normal zero-shot is better than semi-zero shot. However, for less similar languages, semi-zero-shot was better, suggesting that there is some language-specific information aiding the performance. This could be evidence for the language-dependence of emotion detection.

Future research, dealing with better datasets and approaches to make the BERT representations language-neutral, should be carried out to corroborate these findings.

Acknowledgements

This research was funded by Research Foundation–Flanders under a Strategic Basic Research fellowship with grant number 3S004019.[5]

[4] https://github.com/Abhijit-2592/spacy-langdetect

[5] https://www.researchportal.be/nl/project/transfer-learning-voor-automatische-emotiedetectie-nederlandstalige-teksten

References

Lisa Feldman Barrett. 2006. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20–46.

Catherine Caldwell-Harris, Ann Kronrod, and Joyce Yang. 2013. Do more, say less: Saying "I love you" in Chinese and American cultures. *Intercultural Pragmatics*, 10(1):41–69.

Rafael A Calvo and Sunghwan Mac Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3):527–543.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hila Gonen, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. 2020. It's not Greek to mBERT: Inducing word-level translations from multilingual BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 45–56, Online. Association for Computational Linguistics.

Catherine L. Harris, Jean Berko Gleason, and Ayşe Ayçiçeği. 2006. When is a first language more emotional? Psychophysiological evidence from bilingual speakers. In Aneta Pavlenko, editor, *Bilingual Minds: Emotional Experience, Expression, and Representation*, pages 257–283. Multilingual Matters.

Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Kaisla Kajava, Emily Öhman, Piao Hui, Jörg Tiedemann, et al. 2020. Emotion preservation in translation: Evaluating datasets for annotation projection. *Proceedings of Digital Humanities in Nordic Countries (DHN 2020)*, pages 38–50.

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. Universal Joy: A data set and results for classifying emotions across languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75, Online. Association for Computational Linguistics.

Robert I Levy. 1984. The emotions in comparative perspective. *Approaches to emotion*, pages 397–412.

Kristen A. Lindquist, Jennifer K. MacCormack, and Holly Shablack. 2015. The role of language in emotion: Predictions from psychological constructionism. *Frontiers in Psychology*, 6.

Batja Mesquita and Phoebe C Ellsworth. 2001. The role of culture in appraisal. In Klaus R Scherer, Angela Schorr, and Tom Johnstone, editors, *Appraisal processes in emotion: Theory, methods, research*, pages 233–248. Oxford University Press.

Batja Mesquita, Nico H Frijda, and Klaus R Scherer. 1997. Culture and emotion. *Handbook of cross-cultural psychology*, 2:255–297.

Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Woodhead Publishing, Sawston, Cambridge.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Gün R Semin, Carien A Görts, Sharda Nandram, and Astrid Semin-Goossens. 2002. Cultural perspectives on the linguistic representation of emotion and emotion events. *Cognition & Emotion*, 16(1):11–28.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Marc Tanti, Lonneke van der Plas, Claudia Borg, and Albert Gatt. 2021. On the language-specificity of multilingual BERT and the impact of fine-tuning. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 214–227, Punta Cana, Dominican Republic. Association for Computational Linguistics.

B.L. Whorf. 1956. *Language, thought, and reality: Selected writings*. Technology Press of Massachusetts Institute of Technology.

Anna Wierzbicka. 1999. *Emotions across languages and cultures: Diversity and universals*. Cambridge University Press.

Jonathan Winawer, Nathan Witthoft, Michael C Frank, Lisa Wu, Alex R Wade, and Lera Boroditsky. 2007. Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19):7780–7785.