Emoji-Based Transfer Learning for Sentiment Tasks

Susann Boy
Saarland University

Dana Ruiter
Saarland University

Dietrich Klakow
Saarland University

{sboy,druiter,dietrich.klakow}@lsv.uni-saarland.de

Abstract

Sentiment tasks such as hate speech detection and sentiment analysis, especially when performed on languages other than English, are often low-resource. In this study, we exploit the emotional information encoded in emojis to enhance the performance on a variety of sentiment tasks. This is done using a transfer learning approach, where the parameters learned by an emoji-based source task are transferred to a sentiment target task. We analyse the efficacy of the transfer under three conditions, i.e. i) the emoji content and ii) label distribution of the target task as well as iii) the difference between monolingually and multilingually learned source tasks. We find i.a. that the transfer is most beneficial if the target task is balanced with high emoji content. Monolingually learned source tasks have the benefit of taking into account the culturally specific use of emojis and gain up to F1 +0.280 over the baseline.

1 Introduction

Many natural language processing (NLP) tasks suffer from a lack of available data. This is especially true for sentiment tasks, such as hate speech (HS) detection, which depend on the availability of manually annotated data. When moving to languages other than English, many sentiment tasks quickly become very low-resourced.

On the other hand, noisy social media content is available in abundance and many sentiment tasks are based on user comments on such platforms. Emojis can be a valuable source for the distant supervision of sentiment tasks, as they correlate with the underlying emotion of a comment. In this study, we aim to exploit the emotional information encoded in emojis to improve the performance on various sentiment tasks using a transfer learning approach from an emoji-based **source task** (ST) to a sentiment **target task** (TT). Previous work has focused on the transfer from predicting single emojis (Felbo et al., 2017) or strictly pre-defined emoji-clusters (Deriu et al., 2016). However, pre-defined emoji clusters do not take into account the culturally diverse usage of emojis (Park et al., 2012; Kaneko et al., 2019). We therefore introduce data-driven supervised and unsupervised emoji clusters and compare these with single emoji prediction tasks. Specifically, we analyze the efficacy of the transfer from a single emoji or (un)supervised emoji cluster prediction ST to a sentiment TT under three conditions, i.e. *i)* low vs. high amount of **emoji content** present in TT, *ii)* balanced vs. unbalanced **label distribution** in TT and *iii)* **monolingually** or **multilingually** learned ST. The first two conditions are based on typical qualities of sentiment corpora, which tend to be unbalanced in their label distribution with varying degrees of emoji content depending on the source of the data. The third condition is relevant for languages for which a TT is low-resource and which might benefit from a multilingually learned ST.

In Section 2 we give an outline of related work, followed by the introduction of our method (Section 3). The experimental setup in Section 4 details the data and models used as well as the (un)supervised clusters generated. In Section 5 we describe our results and conclude in Section 6.

2 Related Work

Emojis have been used as a type of distant supervision using pre-defined emotion classes based on psychological models (Suttles and Ide, 2013), binary (*positive/negative*) classes (Deriu et al., 2016) or a set of single emojis (Felbo et al., 2017). However, such pre-defined emoji classes often do not account for the culturally diverse use of emojis (Park et al., 2012; Kaneko et al., 2019). In contrast, our

Proceedings of the 16th Conference of the European Chapter of the Associationfor Computational Linguistics: Student Research Workshop, pages 103–110
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

work does not pre-define the emotion classes found in emojis and instead learns these classes, or clusters, from the data itself. While our and the above approaches focus on exploiting emojis as additional labelled data, e.g. in a transfer setting, emoji embeddings (Eisner et al., 2016) have been used as additional features in downstream tasks such as sarcasm detection (Subramanian et al., 2019).

Transfer learning has recently been driven by transformer-based (Vaswani et al., 2017) language models (LM) such as BERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020). When learning a source task on these models, the representations in the encoder change to become informative to the task at hand. In a parameter transfer setting, a new but related target task then profits from the learned representations in the encoder. Transfer learning has been applied to sentiment analysis (SA) using parameter transfer methods such as pre-trained sentiment embeddings (Dong and de Melo, 2018) or machine translation-based context vectors (McCann et al., 2017). Our approach forms part of the parameter transfer approach, as we use encoder representations learned using emoji-based source tasks and transfer these to sentiment target tasks.

Hate speech classification and **sentiment analysis** have in recent years been the object of many shared tasks (Rosenthal et al., 2017; Wiegand, 2018; Basile et al., 2019; Mandl et al., 2019; Ogrodniczuk and Łukasz Kobyliński, 2019). Classification models for these tasks often rely on feature engineering and statistical methods such as naive-bayes (Saleem et al., 2016), logistic regression over subwords (Waseem and Hovy, 2016) or neural approaches including convolutional neural networks (Park and Fung, 2017) or, as in our case, the representations of large LMs (Yang et al., 2019).

3 Method: Emoji-Prediction

For our parameter transfer, we rely on a single transformer-based LM which is shared among different tasks. A sequence $x \in X$ is featurized by reading it into the encoder of the LM and retrieving its last hidden state. A linear layer is then used as a predictive function $f : X \to Y$ to predict labels $y \in Y$. A task $\mathcal{T} = \{Y, f(x)\}$ is then a set of labels Y and the predictive function f over the instances in X.

We follow a **transfer learning** approach, where source task $\mathcal{T}_S$ is an emoji-based classification task, i.e. given a sequence, predict the emoji (class) that it originally contained. Target task $\mathcal{T}_T$ is a downstream task such as SA or HS (Section 4.1). Each task has its own set of instances X, labels Y and predictive function f, while the feature-generating LM stays the same. The error of predictor f is back-propagated to the LM, which allows us to transfer learned parameters from $\mathcal{T}_S$ to $\mathcal{T}_T$.

3.1 Source Tasks (ST)

We focus on 5 different emoji-based STs, that can be divided into two types, emoji prediction (EP) and emoji cluster prediction. To sample emojis for EP or create clusters, we rely on a large collection of user generated comments. **EP** is a multi-class prediction task over the 64 most common emojis identified in the collection of comments. Concretely, given a tweet with all emojis removed, the classifier has to predict which of the 64 emojis was originally contained within it.

The **emoji cluster prediction** tasks can be supervised (PMI-{Target,Swear}) or unsupervised (KMeans-{2,3}). In this case the task is simplified: Given a tweet with all emojis removed, predict the cluster to which the emoji originally contained in the tweet belonged.

Unsupervised Clusters In order to account for the cultural differences in the use of emojis, we learn emoji clusters directly from the user generated data. We generate 50-dimensional vector representations over the tokens in the collection of user comments using the continuous bag of words (Mikolov et al., 2013) approach. We then perform k-means clustering with 6 target clusters on the representations of emojis that occurred ≥ 1000 times. These clusters are manually merged into 2 (*positive/negative*) and 3 (*positive/negative/neutral*) clusters to create the binary **KMeans-2** and ternary **KMeans-3** emoji cluster prediction STs respectively. Below a comment to be classified as *positive* according to the KMeans-{2,3} tasks, as it originally contained an emoji that belonged to the *positive* cluster:

So beautiful and great advice →*positive*

Supervised Clusters As an alternative to the completely unsupervised clusters, we exploit the mutual information between emojis and swear words as a type of distant supervision for HS tasks. We calculate the pointwise mutual information (PMI) between comments in our collection of user content (not) containing slurs and the emojis that appear. An emoji is in the slur cluster if its PMI is larger to comments containing swearwords, otherwise it is in the neutral cluster. **PMI-Swear** is then a binary classification task based on the resulting slur/neutral emoji clusters.

While the unsupervised emoji cluster prediction STs and PMI-Swear are source-oriented, i.e. learned on user generated content, we also explore target-oriented clusters that rely on the shared information between emojis and the labels in each of the TTs. Concretely, we calculate the PMI between the label of an instance in the respective TT training data and the emojis it contains. The emoji is placed into the cluster of the label to which its PMI value is largest. **PMI-Target** is the ST based on these target-oriented emoji clusters.

3.2 Target Tasks (TT)

Once the classifier has been fully trained on the ST, and thus has adapted the underlying LMs representations to fit the ST at hand, we discard it and train a new classifier on top of the enriched LM to predict the TT. We evaluate this transfer from the various STs on two main categories of TTs, namely Hate Speech Detection and Sentiment Analysis. Given a user generated comment, **Hate Speech** Detection is the task of classifying the comment as either *hate* or *none*. Note, however, that concrete label names (e.g. *offense*, *hate*, *harmful*) may differ across specific HS tasks.

While HS in our case is a binary classification task, **Sentiment Analysis** is a ternary classification task which takes as input a user generated comment and classifies it as either *positive*, *neutral* or *negative*. In the following an example from the Sentiment Analysis in Twitter (Rosenthal et al., 2017) task:

> *Finally starting the 5th season of Dexter.*
> *See ya later, weekend! →positive*

Both HS and SA are sentiment-based tasks, e.g. *hate* towards a group of people or *positive* sentiment towards a product etc. We therefore take these two types of tasks to have the potential to benefit from the emotion information encoded in emojis. In the following sections we explore the conditions under which the transfer from an emoji-based ST to a sentiment-based TT is beneficial for the TT.

4 Experimental Setup

We describe the data used for the STs and TTs respectively (Section 4.1), followed by the specifications of the encoding LM (Section 4.2) and the emoji cluster creation (Section 4.3).

Corpus	# Tweets		# Emojis
	Train	Test	
Target Tasks (TT)			
HS-DE	1158/2439	970/2061	853 (7.2%)
SA-DE	1346/900/3676	83/49/197	166 (2%)
HS-ES	1857/2643	660/940	957 (14.5%)
SA-EN	18481/7551/21542	2375/3972/5937	1211 (1.9%)
SA-AR	653/1022/1336	1514/2222/2364	2126 (22.5%)
HS-PL	812/8726	134/866	1733 (13.7%)
Source Tasks (ST)			
TW-DE	16M	–	3M (10%)
TW-EN	323M	–	82M (17%)
TW-ES	320M	–	43M (9%)
TW-PL	7M	–	1M (12%)
TW-AR	183M	–	56M (20%)

Table 1: Number of train, test (for TT) and collected (for ST) tweets as well as number of (non-unique) emojis contained in each corpus. Percentage of training tweets containing emojis in brackets. TTs with label distribution for HS (*hate/none*) and SA (*positive/negative/neutral*) tasks.

4.1 Data

Source Tasks We use a collection[1] of tweets that has been collected from the Twitter stream between 2011 and 2019 as our corpus needed to sample emojis and create emoji clusters for the STs. We perform language identification using the `polyglot`[2] library over the tweets to create a corpus for German, English, Spanish, Polish and Arabic (TW-{DE,EN,ES,PL,AR}) respectively.

To automatically identify swear words for PMI-Swear, we use a German and a multilingual swear word collection, namely `WoltLab`[3] and `Hatebase`[4]. In total, we collected 785 slurs for German, and 1531, 140, 306, 79 for English, Spanish, Polish and Arabic respectively.

Target Tasks We work with 6 target tasks in total, 3 HS and 3 SA tasks, taking into account their emoji content, class (im)balance and language.

For German, we use GermEval 2018 (Wiegand, 2018) Task 1 (*offense/other*) (HS-DE) and SB10k (Cieliebak et al., 2017) (*positive/negative/neutral*) (SA-DE). For English, we use Sentiment Analysis in Twitter (Rosenthal et al., 2017) (*positive/negative/neutral*) (SA-EN). Sentiment Anal-

[1] www.archive.org/details/twitterstream
[2] www.github.com/aboSamoor/polyglot
[3] www.woltlab.com/attachment/3615-schimpfwortliste-txt/
[4] www.hatebase.org/

ysis in Twitter is also used for Arabic (SA-AR). For Spanish we use HatEval (Basile et al., 2019) (*hate*/*none*) (HS-ES) and for Polish, we use PolEval (Ogrodniczuk and Łukasz Kobyliński, 2019) Task 6 (*harmful*/*none*) (HS-PL). For all of the above, we use the original train/test splits. While the HA tasks have different label names, we normalize these to be *hate*/*none* across all tasks. For all SA, the labels to be predicted are *positive*/*negative*/*neutral*.

In Table 1, we report the label distribution, *hate*/*none* for HS and *positive*/*negative*/*neutral* for SA, across all TT training and test sets, as well as ST Twitter corpora sizes. For both ST and TT corpora, we also report the percentage as well as total number of tweets containing emojis.

Preprocessing All data sets undergo the same preprocessing. Tweets are tokenized using the NLTK (Bird and Loper, 2004) `TweetTokenizer` and user mentions, retweets and punctuation are removed. Repeated characters are shortened. We use token frequencies to determine the standard orthography of a word (e.g. *coooool* → *cool* instead of *col*).

4.2 Model Specifications

For the monolingual (German) experiments, we use the German BERT[5] (BERT-DE) and for multilingual experiments we use `Bert-Base-Multilingual-Cased` (BERT-M) as the LM to encode the tweets. We base our code[6] on the `simpletransformers`[7] sequence classification implementations of the above models. Each classification task is trained for a maximum of 10 epochs using early stopping over the validation accuracy with $\delta = 0.01$ and patience 3. Training was performed on a single Titan-X GPU, which took between 1 and 6 hours depending on the data size. We evaluate the resulting classifiers using the Macro F1 measure.

4.3 Clusters

We describe the creation of the emoji clusters used for the emoji cluster STs.

Unsupervised The unsupervised clusters (Section 3) were trained on TW-DE and the concatenation of TW-{DE,EN,ES,PL,AR} for the mono- and multilingual experiments respectively. In both cases, this yielded clusters that can be manually categorized as *happy, love, fun, nature, unhappy, other* (Figure 1). For KMeans-3, {*happy, fun, love*} were merged to *positive*, {*other, nature*} to *neutral* and {*unhappy*} was used as the *negative* class. For KMeans-2, the *neutral* class is ignored.

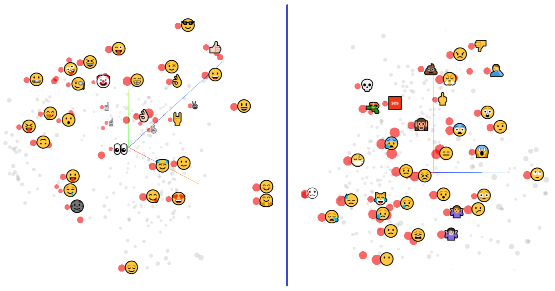

Figure 1: *Happy* (left) and *unhappy* (right) emoji clusters obtained by KMeans on TW-DE.

Supervised The PMI-Target clusters are trained on the respective TT training data. The slur lists are used to identify the slurs in the twitter corpora. PMI-Swear is then trained on TW-DE and the concatenation of TW-{DE,EN,ES,PL,AR} for the mono- and multilingual experiments respectively.

5 Results

We train each model over 10 seeded runs and report the averaged Macro F1 with standard error (Figure 2). For each TT, we train a **baseline**, which is the same pre-trained BERT-{DE,M} model that is now fine-tuned directly on the TT classification task at hand, without prior training on the ST. We compare these baselines with those models that have undergone a transfer from ST to TT. We use the term *equivalent* to signify that two models lie within each others error bounds.

5.1 Condition 1: Emoji Content

We evaluate the effect that STs have on TTs with different amounts of emoji content. We focus on the TTs with the lowest and highest amount of emoji content, namely SA-EN (1.9% emoji content) and SA-AR (22.5%). This is the multilingual case. For the monolingual case, we evaluate the effect on SA-DE (2%) and HS-DE (7.2%). All of these TTs are unbalanced, i.e. the minority class makes up 15.2–32.2% of the training data.

The **monolingual**, low emoji content SA-DE task does not profit from the transfer. Rather, the

[5] www.deepset.ai/german-bert

[6] https://github.com/uds-lsv/emoji-transfer

[7] www.github.com/ThilinaRajapakse/simpletransformers

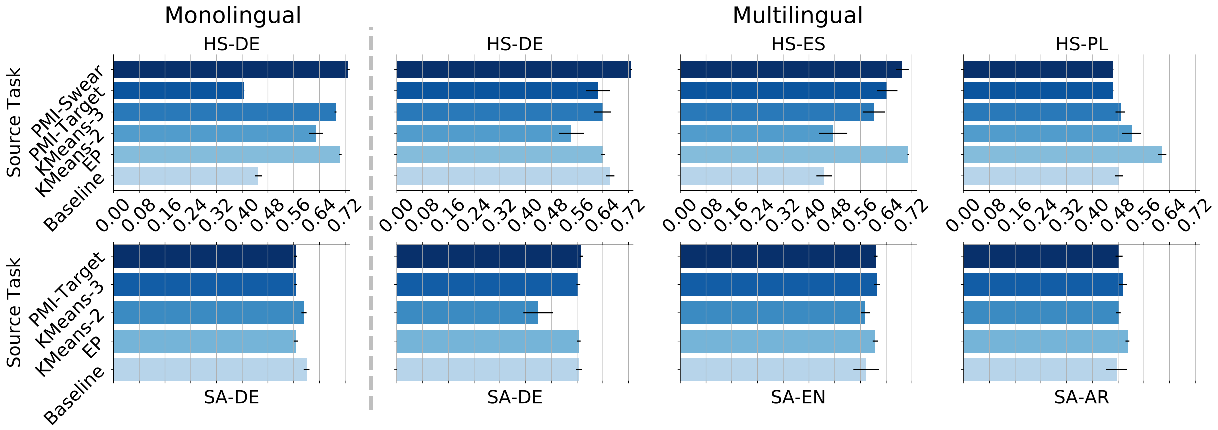

Figure 2: Macro F1 of the HS and SA target tasks transferred from monolingual (left) and multilingual (right) STs.

training on most STs leads to a slight drop in F1-Macro compared to the baseline (F1 0.600). On the other hand, high emoji content HS-DE greatly benefits from the transfer, with PMI-Swear (F1 0.730) being especially beneficial for the performance on the TT, yielding a gain of F1 +0.280 over the baseline. This shows that the shared information in emojis and slurs is relevant to the HS task at hand. Also beneficial are EP (F1 0.705), and the unsupervised KMeans-3 (F1 0.690) and KMeans-2 (F1 0.629) cluster prediction tasks. Only the supervised PMI-Target (F1 0.405) does no seem to be beneficial for the performance on the TT, leading to a drop in performance, which is due to the unbalanced nature of the TT (Section 5.2).

The **multilingual** case shows a slightly mixed trend. Low emoji content SA-EN does not benefit from the transfer, but unlike in the monolingual setting, it is not harmed by it either. All STs lead to a TT performance that is equivalent to the baseline (F1 0.578). High emoji content SA-AR only barely profits from the transfer, with EP (F1 0.509) leading to a small gain of F1 (+0.034) over the baseline (F1 0.475), while all other STs lead to an equivalent performance to the baseline. The overall trend is similar to the monolingual case but the positive and negative effects are dimmed down, which may be due to the multilingual aspect (Section 5.3).

The **general trend** shows that a decent amount of emoji content in the TT training data is crucial for the transfer to be beneficial.

5.2 Condition 2: Label Distribution

To analyze the effect that the STs have on differently (un)balanced TTs, we focus on HS-PL (the minority class makes up 8.5% of training data) and HS-ES (41.3%), as they are the two most (un)balanced TTs, while being comparable in terms of emoji content (13.7% and 14.5% respectively).

For **unbalanced** HS-PL, EP (F1 0.617) and unsupervised KMeans-2 (F1 0.522) lead to an improvement of F1 +0.134 and F1 +0.039 over the baseline, respectively. All other STs are equivalent to the baseline. **Balanced** HS-ES benefits from all TTs, with EP (F1 0.708) leading to a gain of F1 +0.261 over the baseline (F1 0.447), followed by PMI-Swear (F1 0.690) and PMI-Target (F1 0.643). The unsupervised clusters are beneficial but less effective, with F1 0.602 and F1 0.475 for KMeans-3 and KMeans-2 respectively, which likely stems from the multilingual aspect (Section 5.3).

PMI-Target performs poorly on unbalanced HS-PL (and HS-DE etc.) due to its use of mutual information between emojis and the TT labels. This leads to it reproducing the class imbalance, making it less effective on unbalanced TTs.

The difference in impact of **PMI-Swear** on HS-PL (none) and HS-ES (and HS-DE) (gain) can be explained by the composition of the ST dataset. TW-PL is the smallest corpus in the multilingual collection of user comments, and this sparsity is further driven by the morphological complexity of Polish, such that the 306 slurs from the Polish slur list only resulted in $65k$ Polish training samples in PMI-Swear, as opposed to 1.8M and 3M for German and Spanish respectively.

Overall, if the label distribution in TT is balanced, the TT easily benefits from the transfer. Otherwise other conditions such as the multilinguality or emoji content become more relevant.

5.3 Condition 3: Multilinguality

We analyze the effectiveness of the transfer in a monolingual and multilingual setting. For this, we focus on the effect that the monolingually and multilingually learned STs have on HS-DE and SA-DE. Both TTs are unbalanced, while HS-DE has a high emoji content and SA-DE has a low emoji content.

The different effects of the emoji-content in HS-DE and SA-DE has been discussed in Section 5.1, showing that in the **monolingual** setting, high emoji content HS-DE benefits from the transfer, while low emoji content SA-DE does not. In the **multilingual** case, we see a similar, but dimmed, trend. SA-DE does not benefit from the transfer, with all TTs leading to an equivalent performance as the baseline (F1 0.566), except KMeans-2 (F1 0.439) which is below the baseline. The STs have a similar performance on HS-DE, being equivalent or below the baseline (F1 0.663). Only PMI-Swear (F1 0.678) is beneficial for the TT performance.

The effect of ST-oriented clusters KMeans-{2,3} was beneficial in the monolingual case (HS-DE), but this benefit is lost in the multilingual setting. This underlines our original idea that ST-oriented unsupervised emoji clusters learned on large amounts of user generated text have the advantage of accounting for **cultural differences** in the usage of emojis. When learned multilingually, this advantage is lost. An example of the culturally diverse use of emojis is ♻, which is rather infrequent in Europe and might be used to point towards the importance of *recycling*. In TW-AR, this emoji is among the top 5 most frequent emojis, and is used to motivate other users to *share* their content.

The **overall trend** thus shows that monolingually learned STs are more beneficial than multilingual STs. However, if the training data of a TT is balanced, this effect is less pronounced.

5.4 Comparison to Benchmark Results

To put the results into a broader perspective, we compare to state-of-the-art (SOTA) models for each of the shared-tasks/datasets that our TTs are based on (Table 2). For two of the **Hate Speech** benchmarks, the performance of our transfer approach is close to the SOTA, namely with a difference of F1 -0.038 (HS-DE) and F1 -0.03 (HS-ES). For HS-PL, we were able to achieve a gain of $+0.031$ over the SOTA. Across all three **Sentiment Analysis** benchmarks, our models are below the SOTA. This indicates that SA, in general, is a more difficult task to our transfer approach than HS, possibly due to its ternary, rather than binary, classification objective. This is another factor causing the transfer to be overall more beneficial for HS rather than SA, next to the unbalanced (SA-{EN,AR}) and low-emoji content (SA-DE) nature of the SA tasks.

TT	Method	F1	SOTA
HS-DE	PMI-Swear (monolingual)	0.730	**0.768**
HS-ES	EP	0.708	**0.730**
HS-PL	EP	**0.617**	0.586
SA-DE	Baseline (monolingual)	0.600	**0.651**
SA-AR	EP	0.509	**0.610**
SA-EN	KMeans-3	0.611	**0.677**

Table 2: Macro F1 comparison of top-scoring transfer method (*F1*) with SOTA results on the different TT test sets. Best scores in **bold**. See (Montani and Schüller, 2018) (HS-DE), (Basile et al., 2019) (HS-ES), (Ogrodniczuk and Łukasz Kobyliński, 2019) (HS-PL), (Cieliebak et al., 2017) (SA-DE) and (Rosenthal et al., 2017) (HS-{AR,EN}) for SOTA method descriptions.

6 Summary

We have evaluated and identified conditions under which the transfer from an emoji-based ST is beneficial for a sentiment TT. In the experiments in Section 5 we observed three major trends, namely i) TTs with high amounts of emoji content benefit more from the transfer, ii) PMI-Target tends to be detrimental to unbalanced TTs and iii) monolingually learned STs tend to perform better than their multilingual counterparts, due to their improved representation of culturally unique emoji usages. The latter underlines the importance of taking into account cultural differences when exploiting the information encoded in emojis.

From these results, we can draw conclusions about the conditions under which a given emoji-based ST is beneficial. Due to the shared information between emojis and slurs, **PMI-Swear** is beneficial to HS tasks when the data that can be generated from the swear word list is decently large. **PMI-Target** is beneficial when the TT is balanced, otherwise it replicates the already existing class imbalance. Unsupervised **KMeans-{2,3}** should be learned monolingually to be beneficial and **EP** is a safe choice for TTs with high emoji content.

Acknowledgments

We want to thank the anonymous reviewers as well as Thomas Kleinbauer for their valuable feedback. The project on which this paper is based is funded by the DFG under funding code WI 4204/3-1.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics.

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

Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurelien Lucchi, Valeria De Luca, and Martin Jaggi. 2016. SwissCheese at SemEval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1124–1128, San Diego, California. Association for Computational Linguistics.

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.

Xin Dong and Gerard de Melo. 2018. A helping hand: Transfer learning for deep sentiment analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2524–2534, Melbourne, Australia. Association for Computational Linguistics.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Daisuke Kaneko, Alexander Toet, Shota Ushiama, Anne-Marie Brouwer, Victor Kallen, and Jan B.F. van Erp. 2019. Emojigrid: A 2d pictorial scale for cross-cultural emotion assessment of negatively and positively valenced food. *Food Research International*, 115:541 – 551.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, volume 30, pages 6294–6305. Curran Associates, Inc.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Joaquín Padilla Montani and Peter Schüller. 2018. Tuwienkbs at germeval 2018: German abusive tweet detection. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 45–50.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2019. *Proceedings of the PolEval 2019 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

Jaram Park, Young Min Baek, and Meeyoung Cha. 2012. Cross-Cultural Comparison of Nonverbal Cues in Emoticons on Twitter: Evidence from Big Data Analysis. *Journal of Communication*, 64(2):333–354.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop* *on semantic evaluation (SemEval-2017)*, pages 502–518.

Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2016. A web of hate: tackling hateful speech in online social spaces. In *First Workshop on text Analytics for Cybersecurity and Online Safety at LREC 2016*.

Jayashree Subramanian, Varun Sridharan, Kai Shu, and Huan Liu. 2019. Exploiting emojis for sarcasm detection. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 70–80. Springer.

Jared Suttles and Nancy Ide. 2013. Distant supervision for emotion classification with discrete binary values. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 121–136. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Michael Wiegand. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. pages 1–10, Wien. Verlag der Österreichischen Akademie der Wissenschaften.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.