Overcoming Rare-Language Discrimination in Multi-Lingual Sentiment Analysis

Jasmin Lampert
Competence Unit Data Science & Artificial Intelligence
Center for Digital Safety & Security
AIT Austrian Institute of Technology
Vienna, Austria

Christoph H. Lampert
Machine Learning and Computer Vision Group
Institute of Science and Technology Austria (IST Austria)
Klosterneuburg, Austria

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9672003

***Abstract*—The digitalization of almost all aspects of our everyday lives has led to unprecedented amounts of data being freely available on the Internet. In particular social media platforms provide rich sources of user-generated data, though typically in unstructured form, and with high diversity, such as written in many different languages.**

Automatically identifying meaningful information in such big data resources and extracting it efficiently is one of the ongoing challenges of our time. A common step for this is *sentiment analysis*, which forms the foundation for tasks such as opinion mining or trend prediction. Unfortunately, publicly available tools for this task are almost exclusively available for English-language texts. Consequently, a large fraction of the Internet users, who do not communicate in English, are ignored in automatized studies, a phenomenon called *rare-language discrimination*.

In this work we propose a technique to overcome this problem by a truly multi-lingual model, which can be trained automatically without linguistic knowledge or even the ability to read the many target languages. The main step is to combine self-annotation, specifically the use of emoticons as a proxy for labels, with multi-lingual sentence representations.

To evaluate our method we curated several large datasets from data obtained via the free Twitter streaming API. The results show that our proposed multi-lingual training is able to achieve sentiment predictions at the same quality level for rare languages as for frequent ones, and in particular clearly better than what mono-lingual training achieves on the same data.[1]

***Index Terms*—sentiment analysis, algorithmic fairness, multi-lingual sentence embeddings, self-annotation, natural language processing, social media**

I. Introduction

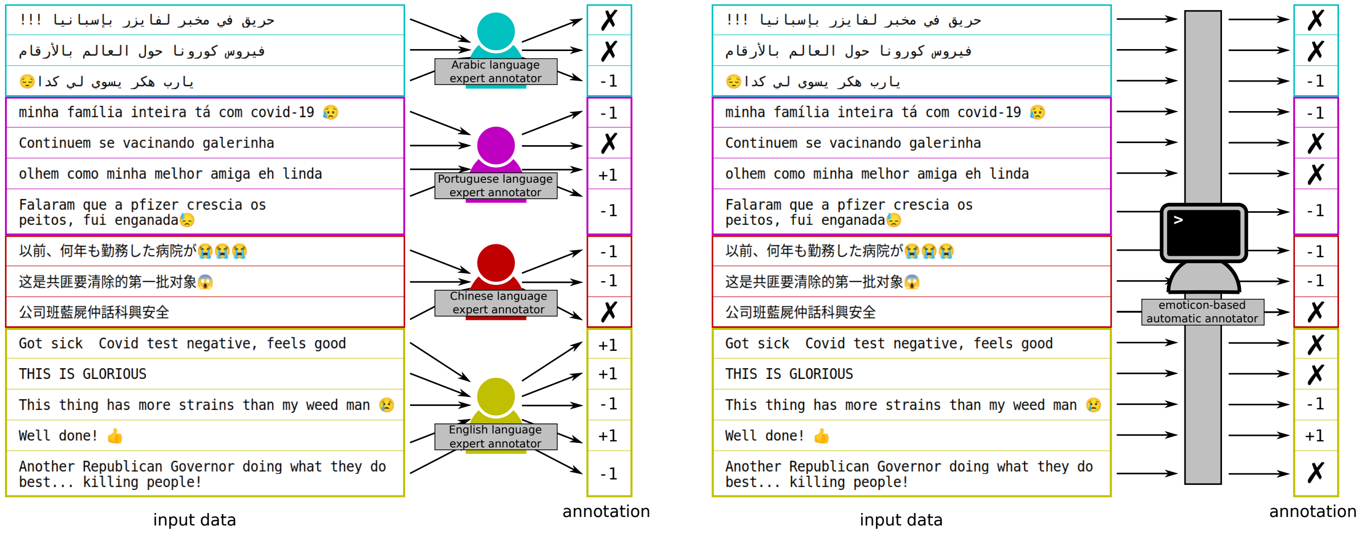
Social media, such as Facebook or Instagram, and microblogging platforms, such as Twitter, offer public access to a rich stream of information, which can readily be exploited using suitable tools for natural language processing (NLP) [1], [2], [3], [4]. For example, sentiment analysis of twitter feeds allow the tracking of trends in real-time, e.g., the popularity of products or of political candidates, as well as for analyzing and predicting election outcomes [5], [6], [7]. Additionally, it has been successfully applied in the domain of crisis and disaster management, e.g., for organizing disaster response and in public health, e.g., for gaining a better understanding of a pandemic disease outbreak [8], [9].

It is often overlooked, however, that such analyses typically come with a strong language bias: while only approximately 5% of the world population has English as their native tongue, almost all publicly available NLP tools, in particular for sentiment analysis, are applicable only to English-language texts. With social media analysis increasingly often being used not only to observe societal trends but also as foundation for decision making, e.g. in politics, the lack of representation of smaller languages is not only a nuisance but can lead to a real-world discrimination against speakers of these languages: if they are invisible to automatic monitoring tools, their interests are not represented.

Historically, the focus on English-language tools can be understood from the fact that, originally, creating sentiment analysis systems required a lot of time of experts with technical knowledge about NLP techniques. Building multilingual systems would have been prohibitively expensive, as the needed time and cost grows essentially linear in the number of languages covered. Concentrating on English as only language made sense, because methodological progress was largely driven by the international academic community, where English is the lingua franca.

In today's time of big data and high performance computing, deep expert knowledge is not required anymore to build powerful NLP tools. Instead, state-of-the-art systems for sentiment analysis are created using *supervised machine learning*, in which the computer autonomously learns the association between words or other textual elements with positive or negative sentiment. The main ingredient for this is a *training set*, i.e. a dataset of natural language texts (inputs), which has been annotated with the information whether they express positive or negative sentiment (ground truth output). Nevertheless, even today's machine-learning-based sentiment analysis tools are available almost exclusively for English, and that is because almost all available datasets contain only English-language texts (with a few and typically smaller exceptions for other big languages, such as Spanish or Chinese [10], [11]). There are two reasons for this: first, the easiest and cheapest way to collect large amount of textual data is to use Internet resources, where English is the most-frequent language, while rare languages constitute only a tiny fraction of all available data. Second, the limiting factor in creating datasets for supervised

[1] Our source code, pretrained models and follow-up works are available at https://github.com/clampert/multilingual-sentiment-analysis

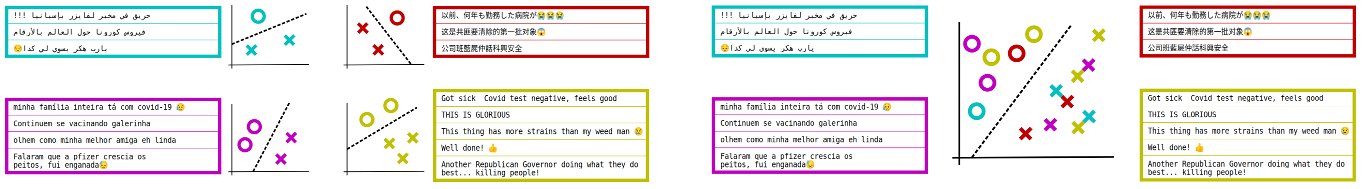

Arabic language expert annotator
Portuguese language expert annotator
Chinese language expert annotator
English language expert annotator
emoticon-based automatic annotator
input data
annotation
✗
-1
+1

حريق في مخزن لفايزر بإسبانيا !!!
فيروس كورونا حول العالم بالأرقام
يارب هكر يسوي لي كدا😔
minha família inteira tá com covid-19 😥
Continuem se vacinando galerinha
olhem como minha melhor amiga eh linda
Falaram que a pfizer crescia os peitos, fui enganada😓
以前、何年も勤務した病院が😭😭😭
这是共匪要清除的第一批对象😱
公司班藍屍仲話科興安全
Got sick Covid test negative, feels good
THIS IS GLORIOUS
This thing has more strains than my weed man 😢
Well done! 👍
Another Republican Governor doing what they do best... killing people!

(a) Traditional (manual) annotation

(b) Self-annotation

Fig. 1: Illustration: (a) In the traditional annotation approach, for each language a human annotator with knowledge of that language is required. (b) In the self-supervised approach, the annotation is generated automatically from emoticons in the text.



(a) Traditional (mono-lingual) sentence embeddings

(b) Multi-lingual sentence embeddings

Fig. 2: Illustration: (a) With traditional (mono-lingual) sentences embeddings, the representations of different languages are inconsistent, and a separate classifier is required for each language. (b) With a multi-lingual embedding, texts of all languages are embedded consistently, and it suffices to train a single classifier.

learning is typically not to collect input data, but to acquire the *ground truth* outputs from human annotators. Having to recruit annotators not only for one but a large number of (potentially rare) languages is expensive and time-consuming, which again means that the effort of creating multi-lingual tools grows with the number of languages covered.

As our main contribution in this work we propose a technique to mitigate rare-language discrimination, which we demonstrate by creating a sentiment analysis system that is truly multi-lingual by construction. Its most important aspect is that the training set is generated fully automatic. Thereby, the required human effort is reduced to a minimal amount, which -in particular- is independent of the number of languages covered. We achieve this task by building on two recent developments in natural language processing: *self-supervised training* and *multi-lingual sentence embeddings*.

A. Method Sketch

Self-supervised training is a way to overcome the need of annotating large data corpora: instead of having a human provide ground truth outputs, the computer computes these itself from contextual data. Fig. 1) illustrates this. Specifically, we suggest to automatically generate annotations for the sentiment analysis task from readily available unlabeled tweets by analyzing the emoticons they contain. Since most tweets do not contain any emoticons, this procedure is not a drop-in replacement for creating a sentiment analysis system in general. However, given a large enough source of input data, the smaller fraction of tweets that do contain suitable emoticons suffices to create a training set, from which a system can be trained that assigns sentiments also to tweets without emoticons.

getting the training data by using emoticons as
a way for analyze the sentiment of the tweets

Besides avoiding the need for manual annotation, a second advantage is that with the widespread use of UTF-8 text encodings, emoticons are encoded consistently between different languages. Therefore, the same self-annotation routines can be applied to data from many different languages, and the actual language of the text does not even have to be known. The result is training data for sentiment classification of as many languages as the original data source contained and the computational cost for this is only proportional to the amount of available data, not to the number of languages covered.

The resulting dataset can be readily used to train a sentiment classification system for any included language. However, More data, more emoticons more labelling that could counter the effects of rare language mitigation. But this requires more data anyways

the results of this would still be a system that exhibits rare language discrimination. The reason is that language distribution for data generated this way is highly imbalanced. For some languages orders of magnitude less training data is available than others. Because classification accuracy typically increases monotonically with the amount of available training data, simply training individual classifiers on this data for each language would still result in an algorithmic bias: classifiers for rare languages would be systematically worse than for common languages, thereby reducing their practical value.

The use of *multi-lingual sentence embeddings* (Fig. 2) for the textual data is our way to overcome this problem. Sentence embeddings are common software modules in NLP that take as input a variable-length piece of text and output a vector representation of fixed dimensionality, which can subsequently be used, e.g., as input to a classifier. The characteristic property of multi-lingual embeddings is that sentences of similar meaning are represented by similar vectors, even if they were written in different languages. As a consequence, after the embedding step has taken place, the difference between languages mostly disappears. More precisely, a classifier trained on text in one language (in its embedded form) still works to some extent on text in another language (embedded in the same way).

Embeddings contains the sentiments of the text also

The combination of both techniques allows us to train for the first time a classifier that is able to assess the sentiment of short texts in any of many different languages, and that works comparably well for all of them, thereby avoiding a bias against rare languages. The technical details are provided in Section II.

B. Related work

It is well-known that data from social media platforms is a rich source of publicly available information, which can be leveraged using natural language processing. For a detailed discussion, see, for example, [4]. Sentiment analysis, e.g. [3], is a particularly popular technique in this area, as it enables a multitude of applications, including some of commercial value, such as the automatic assessment of the popularity of products [5], politicians [12] or societal trends [13] over time.

Systems for sentiment analysis are typically either rule-based [14] or machine-learning-based [15]. In both cases, however, they are typically mono-lingual, i.e. applicable only to a single language, and most often this is English.

In this work, we use two recent trends in natural language processing to overcome the rare language problem. *Self-annotation* has been popularized, e.g. for the tasks of learning word representations [16] and natural language generation [17]. It was not usable for sentiment analysis, though, until the recent rise of social media texts, which contain not only words but also expressions of emotions, i.e. emoticons. One of the first works to observe this was [18], though only in a mono-lingual setting (here English), and looking only for smileys in ASCII-text form. Later, emoticons in UTF-8-format and in other languages than English were explored, such as Greek in [19][2]

Multi-lingual text embeddings are created by combining standard representation learning approaches with constraints that words or sentences, which are translations of each other, should be mapped to similar representations. This requires paired data, i.e. textual documents that we know are translations of each other. This kind of data became widespread for public use, on the one hand, with the growing popularity of international variants of Wikipedia and, on the other hand, with the widespread distribution of DVDs, which contain subtitle tracks in many different languages [20], [21]. Such representations were able to handle two or a few more languages, but not tens or a hundred of them. The latter only became possible quite recently through the integration of incremental training via knowledge distillation [22].

II. Method

A. Dataset and self-annotation

For our experiments we use a 5.6 TB dataset containing approximately 740,000,000 multi-lingual tweets. It was collected between March 2020 and June 2021 using Twitter's Streaming API[3] with varying COVID-19-related keywords. In this data we identify all tweets that contain either positive or negative emoticons according to Fig. 3. Tweets that contain both positive and negative emoticons are excluded. Each such tweet is assigned a positive or negative label depending on the occurring emoticons. Overall, we obtain an annotated dataset with 9,481,337 tweets, which is far larger than annotated datasets typically used in the literature [23], [24], [25]. The label proportions are roughly balanced (56% positive to 44% negative sentiment labels). Table I shows the occurring languages (abbreviation and full names) as well as their frequencies. As one can see, the frequency of the different languages is approximately Zipf-distributed [26], with a small number of languages that occur very frequently, but also a large number of languages that occur very rarely. Such rare languages are a problem for learning-based methods, since standard classification techniques usually require thousands of examples to achieve acceptable performance.

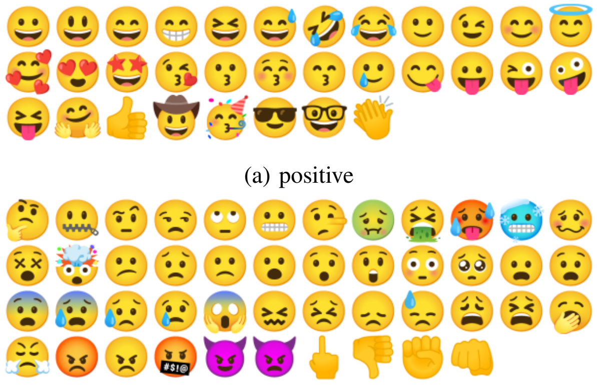
Given that most languages in our dataset do not reach this threshold, one might be tempted to suggest to collect even more data. This, however, is not a practical way to solve the rare-languages problem. For example, to ensure that at least 1000 example sentences are available for each of the occurring languages (except Tibetan), a dataset approximately 50–100 times larger than ours (i.e. 500 million to 1 billion examples) would be required. This, however, would likely cause new, even rarer, languages appear, similar to how Tibetan occurs in our dataset for the first (and only) time in data from March 2021, more than a year after the data collection process started.

[2] Despite the work's title, the system described in [19] is actually mono-lingual by our terminology, as it only uses Greek texts as input.

[3] https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data

TABLE I: Language distribution in dataset. Horizontal lines separate logarithmically-sized bins of fraction.

Language	Number of Tweets	Fraction
`en` (English)	5,006,797	0.5280690
`es` (Spanish)	1,182,889	0.1247600
`fr` (French)	633,958	0.0668638
`de` (German)	473,464	0.0499364
`in` (Indonesian)	381,842	0.0402730
`hi` (Hindi)	267,414	0.0282043
`tr` (Turkish)	254,928	0.0268873
`pt` (Portuguese)	245,273	0.0258690
`tl` (Tagalog)	192,083	0.0202591
`it` (Italian)	153,170	0.0161549
`nl` (Dutch)	136,424	0.0143887
`ja` (Japanese)	82,347	0.0086852
`ar` (Arabic)	60,657	0.0063975
`th` (Thai)	55,694	0.0058741
`ca` (Catalan)	41,667	0.0043946
`pl` (Polish)	40,257	0.0042459
`et` (Estonian)	35,959	0.0037926
`ta` (Tamil)	27,181	0.0028668
`ht` (Haitian)	21,637	0.0022821
`ur` (Urdu)	15,072	0.0015896
`da` (Danish)	14,015	0.0014782
`ru` (Russian)	13,464	0.0014200
`sv` (Swedish)	13,284	0.0014011
`el` (Greek)	13,260	0.0013985
`mr` (Marathi)	11,796	0.0012441
`fi` (Finnish)	11,642	0.0012279
`zh` (Chinese)	10,811	0.0011402
`cs` (Czech)	8,926	0.0009414
`fa` (Persian)	7,066	0.0007453
`ro` (Romanian)	6,999	0.0007382
`ne` (Nepali)	6,647	0.0007011
`sl` (Slovenian)	5,531	0.0005834
`eu` (Basque)	4,156	0.0004383
`te` (Telugu)	4,034	0.0004255
`no` (Norwegian)	4,022	0.0004242
`gu` (Gujarati)	3,744	0.0003949
`cy` (Welsh)	3,278	0.0003457
`lt` (Lithuanian)	3,071	0.0003239
`ml` (Malayalam)	2,950	0.0003111
`kn` (Kannada)	2,887	0.0003045
`lv` (Latvian)	2,841	0.0002996
`si` (Sinhala)	2,616	0.0002759
`hu` (Hungarian)	2,606	0.0002749
`bn` (Bengali)	2,104	0.0002219
`ko` (Korean)	2,064	0.0002177
`vi` (Vietnamese)	1,824	0.0001924
`or` (Oriya)	1,236	0.0001304
`is` (Icelandic)	1,070	0.0001129
`sr` (Serbian)	965	0.0001018
`uk` (Ukrainian)	952	0.0001004
`iw` (Hebrew)	567	0.0000598
`bg` (Bulgarian)	519	0.0000547
`pa` (Panjabi)	328	0.0000346
`am` (Amharic)	292	0.0000308
`sd` (Sindhi)	188	0.0000198
`hy` (Armenian)	185	0.0000195
`ckb` (Central Kurdish)	156	0.0000165
`ps` (Pushto)	135	0.0000142
`dv` (Divehi)	134	0.0000141
`lo` (Lao)	120	0.0000127
`my` (Burmese)	88	0.0000093
`km` (Central Khmer)	34	0.0000036
`ka` (Georgian)	16	0.0000017
`bo` (Tibetan)	1	0.0000001



(a) positive

(b) negative

Fig. 3: Emoticons used for self-annotation of (a) positive and (b) negative sentiments.

B. Multi-Lingual Sentence Embeddings

For each tweet in the dataset, we first remove the detected emoticons from the text. We then compute a 768-dimensional vector embedding using the multi-lingual sentence transformer model *stsb-xlm-r-multi-lingual*[4] [22]. This representation is trained on 53 languages such that sentences of different languages are assigned a similar vector representation if they have similar semantics. While some of the languages in our dataset, e.g. Icelandic, are not part of the training languages, we found the representation nevertheless to be useful for them. Presumably this is because some languages used for training are sufficiently related, such as Norwegian for Icelandic.

Instead of working with a single very large dataset we split our data into 16 disjoint datasets, one per month. In this way we are able to measure also the variability of the method with different training sets and assess the statistical significance of our results. Each subset contains between 140,000 and 1,400,000 examples, which we partition into equally-sized parts for model training and evaluation.

C. Sentiment Classification

To obtain a sentiment classifier we train a Logistic Regression model, i.e. a linear classifier with a sigmoid activation, where the latter ensures that the predicted values can be interpreted as probability estimates. In preliminary experiments we found that other classifier architectures, in particular multi-layer neural networks, did not lead to improved performance. Presumably, this is because the sentence embeddings are the result of a trained deep network already, such that the subsequent classification task itself does not require additional non-linearity.

For each month we train the model parameters using the Adam optimizer [27] of tensorflow's keras framework for 10 epochs. The batchsize is 100 and the learning rate is left at its default value of 0.001. For the proposed multi-lingual training we use the data of all languages jointly to form a single large

[4] Publicly available at https://www.sbert.net/docs/pretrained_models.html

They found that the resulting multi layered neural nets didnt gave an improved performance and this is because the embedding already contains a large neural net and a lot of non-linear layer.
They also found that this wont help as we already did it !! [Interesting observation]

TABLE II: Classification accuracy (ACC) and area under the ROC curve (AUC) for multi-lingual training (proposed) and for mono-lingual training (baseline).

	overall ACC	overall AUC
multi-lingual	0.691 ± 0.004	0.755 ± 0.005
mono-lingual	0.696 ± 0.004	0.760 ± 0.005

(a) Overall ACC and AUC (each example has equal weight)

	per-language ACC	per-language AUC
multi-lingual	0.670 ± 0.006	0.691 ± 0.008
mono-lingual	0.642 ± 0.011	0.622 ± 0.012

(b) Per-language ACC and AUC (each language has equal weight)

optimization problem. This makes sense, because the tweets of all languages are represented in a semantically consistent way in a shared feature space. Without such a shared representation, a joint classifier would not make sense, and separate classifiers would have to be trained for each language.

As a baseline we implement such a mono-lingual approach. Making use of the fact that the language in which the tweets are written is part of the available meta-data, we split the training and test sets accordingly into mono-lingual ones. We then use the same sentence embedding and training procedure as above to learn a separate mono-lingual classifiers for each language that occurs with at least 10 samples in any of the training sets (this is all except Central Khmer, Georgian and Tibetan). This baseline measures in how far the fact, that rare languages have small training sets, influences the quality of the learned classifiers.

A hypothetical further baseline could be to run automatic translation on all tweets in order to apply existing English-language sentiment classifiers. This, however, is impractical, because multi-lingual translation services typically charge at least 0.01 USD per API call, so translating all tweets in our dataset would incur costs of nearly 100,000 USD. Furthermore, a system based on external automatic translation services would have much higher latency than a single neural network call, and also faces potential privacy issues.

We evaluate all resulting models on the evaluation part of the data, always computing the *accuracy (ACC)* as well as the *area under the ROC-curve (AUC)*. For comparisons we prefer the latter, since it is invariant to the label proportions and therefore can be compared on an absolute scale (1 indicating perfect performance; 0.5 corresponding to chance level).

We are particularly interested in the effect of multi-lingual versus mono-lingual training on rare languages. To be able to study this, we also perform the evaluation separately for each language.

III. Results

A. Quantitative Results

Table II reports the prediction quality of the resulting classifiers as the mean and standard error across all datasets. From Table IIa, which shows the overall average results across the whole evaluation set (i.e. each sample has equal weight when averaging), one might get the impression that multi-lingual and mono-lingual training lead to similar outcomes, as they result in almost identical overall accuracy and AUC values.

This is a misconception, though, as can be seen from Table IIb: it reports the average per-language quality (i.e. each language has equal weight), and, with respect to this measure, multi-lingual training achieves clearly better results. The reason for the discrepancy is the highly imbalanced distribution of the language frequencies (see Table I). Frequent languages heavily dominate the overall average, while differences in the prediction quality for rare languages have hardly any influence on the overall measures. Consequently, judging methods simply based on their overall prediction quality can hide an existing bias against rare languages.

Figures 4 and 5 shed light on where the differences between multi-lingual and mono-lingual training lie. The top plot in each of them shows the average AUC and accuracy values separately for each language, as well as the standard error of the average across the multiple datasets. The bottom plot makes the differences between the multi-lingual and the mono-lingual system explicit. The languages are sorted by decreasing frequency in the data.

One can see that the mono-lingual classifier and the multi-lingual classifier perform comparably on languages that are very frequent (in particular English, which is present with over 150,000 examples on average in each training set) to medium-frequent (such as Catalan, for which each training set contains approximately 1,300 examples on average). However, for even less frequent languages, differences between the methods emerge. The quality of the mono-lingual classifier deteriorates, until for very rare languages (such as Oriya, which on average occurs less than 40 times in the training sets) its accuracy and AUC are often near chance level. This rare-language bias is easy to spot visually, with the curves of average values decreasing from left to right. In contrast, the prediction quality of multi-lingual training is almost equally high across all languages, which is visible as a curve of average values that remains nearly constant. In particular, the prediction quality for rare languages is clearly higher in our proposed multi-lingual training setup than what would be achieved with a mere mono-lingual training.

A Wilcoxon signed-rank test [28] confirms that the differences between the multi-lingual system and the mono-lingual ones are statistically significant at a 10^{-9} and 10^{-3} level for the AUC and the accuracy, respectively.

A second visible property of the curves is that the error bars increase from left to right. This is not just a sign of an unstable training process (as in fact, the multi-lingual training is very stable), but rather a consequence of the small size of the test set for rare languages.

B. Qualitative Results

Table III shows some tweets with highly positive or highly negative sentiment predictions from different languages. In

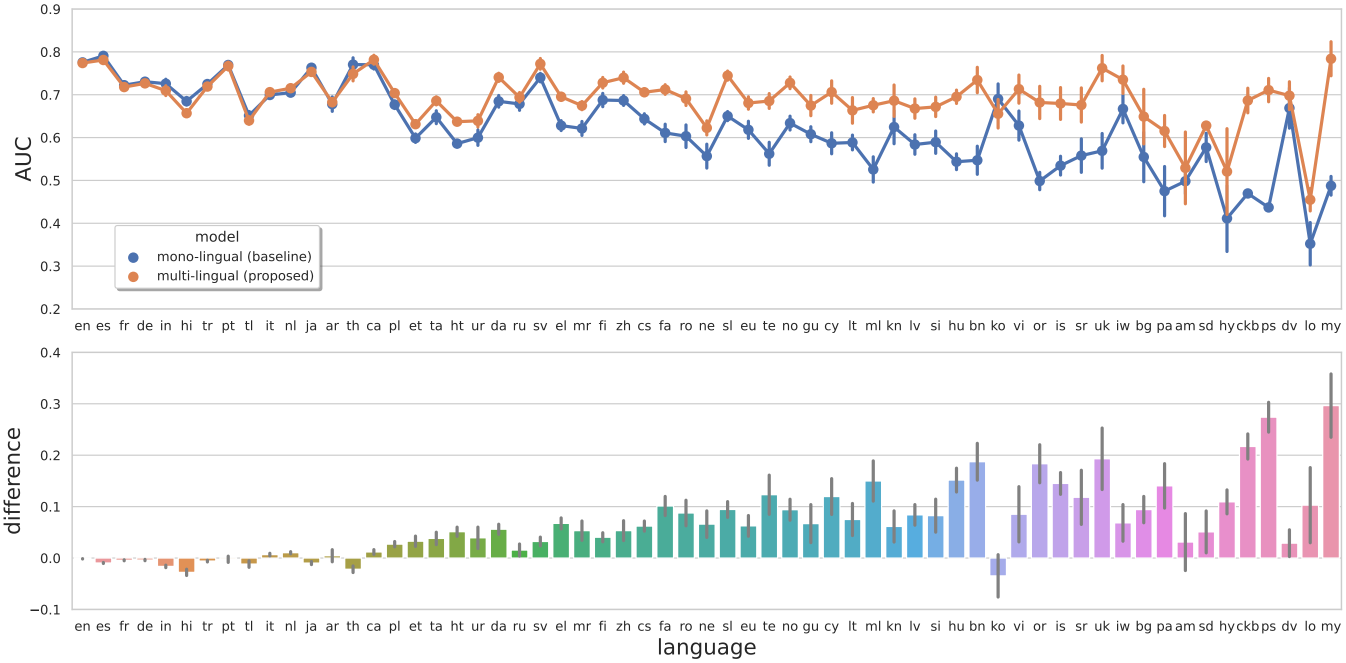


Fig. 4: Top: Per-language evaluation of sentiment classifier quality (area under the ROC curve: AUC) using mono-lingual (baseline) or multi-lingual (proposed) training. Languages on the x-axis are sorted according to language frequency. Bottom: Differences between both training methods.

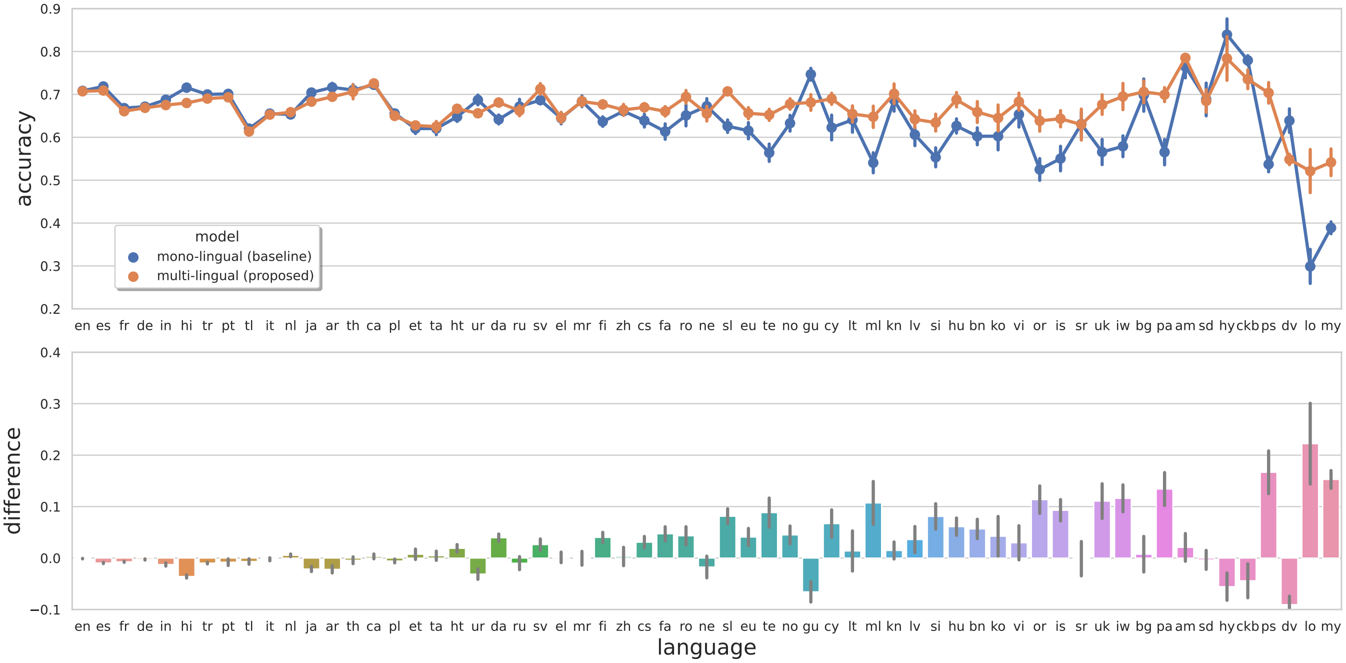


Fig. 5: Top: Per-language evaluation of sentiment classifier quality (classification accuracy) using mono-lingual (baseline) or multi-lingual (proposed) training. Languages on the x-axis are sorted according to language frequency. Bottom: Differences between both training methods.

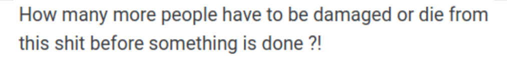
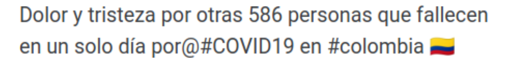
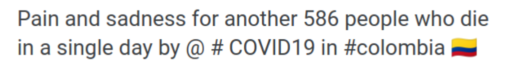
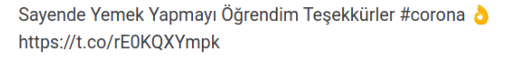
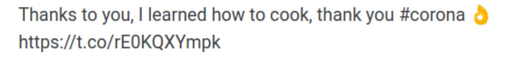
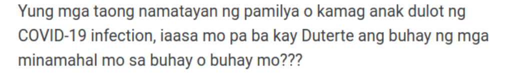
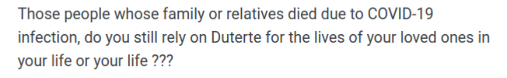
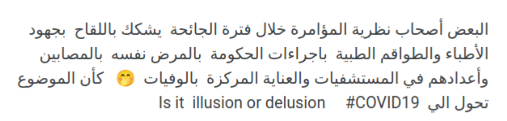
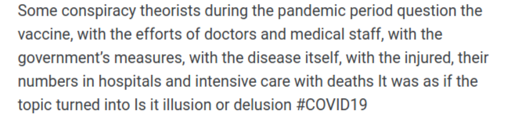
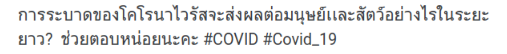
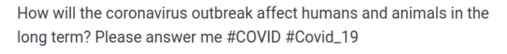
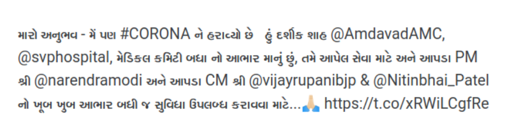
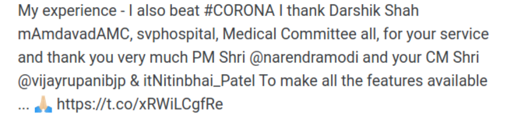
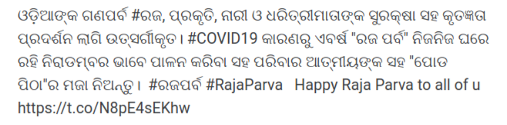
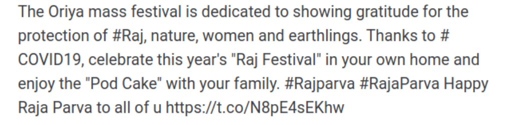
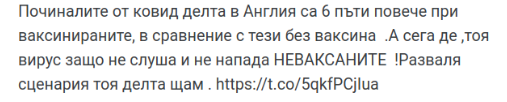
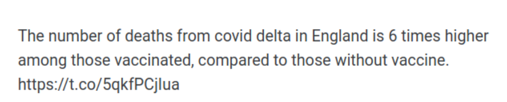
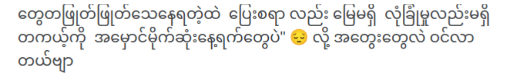
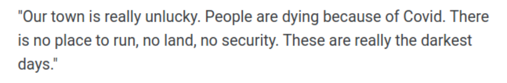
Language	Original tweet text	English translation	Sentiment
en (English)	 How many more people have to be damaged or die from this shit before something is done ?!		negative
es (Spanish)	 Dolor y tristeza por otras 586 personas que fallecen en un solo día por@#COVID19 en #colombia 🇨🇴	 Pain and sadness for another 586 people who die in a single day by @ # COVID19 in #colombia 🇨🇴	negative
tr (Turkish)	 Sayende Yemek Yapmayı Öğrendim Teşekkürler #corona 👌 https://t.co/rE0KQXYmpk	 Thanks to you, I learned how to cook, thank you #corona 👌 https://t.co/rE0KQXYmpk	positive
tl (Tagalog)	 Yung mga taong namatayan ng pamilya o kamag anak dulot ng COVID-19 infection, iaasa mo pa ba kay Duterte ang buhay ng mga minamahal mo sa buhay o buhay mo???	 Those people whose family or relatives died due to COVID-19 infection, do you still rely on Duterte for the lives of your loved ones in your life or your life ???	negative
ar (Arabic)	 البعض أصحاب نظرية المؤامرة خلال فترة الجائحة يشكك باللقاح بجهود الأطباء والطواقم الطبية باجراءات الحكومة بالمرض نفسه بالمصابين وأعدادهم في المستشفيات والعناية المركزة بالوفيات 🤭 كأن الموضوع تحول الي #COVID19 Is it illusion or delusion	 Some conspiracy theorists during the pandemic period question the vaccine, with the efforts of doctors and medical staff, with the government's measures, with the disease itself, with the injured, their numbers in hospitals and intensive care with deaths It was as if the topic turned into Is it illusion or delusion #COVID19	negative
th (Thai)	 การระบาดของโคโรนาไวรัสจะส่งผลต่อมนุษย์และสัตว์อย่างไรในระยะยาว? ช่วยตอบหน่อยนะคะ #COVID #Covid_19	 How will the coronavirus outbreak affect humans and animals in the long term? Please answer me #COVID #Covid_19	negative
gu (Gujarati)	 મારો અનુભવ - મેં પણ #CORONA ને હરાવ્યો છે હું દર્શીક શાહ @AmdavadAMC, @svphospital, મેડિકલ કમિટી બધા નો આભાર માનું છું, તમે આપેલ સેવા માટે અને આપડા PM શ્રી @narendramodi અને આપડા CM શ્રી @vijayrupanibjp & @Nitinbhai_Patel નો ખૂબ ખુબ આભાર બધી જ સુવિધા ઉપલબ્ધ કરાવવા માટે...🙏 https://t.co/xRWiLCgfRe	 My experience - I also beat #CORONA I thank Darshik Shah mAmdavadAMC, svphospital, Medical Committee all, for your service and thank you very much PM Shri @narendramodi and your CM Shri @vijayrupanibjp & itNitinbhai_Patel To make all the features available ... 🙏 https://t.co/xRWiLCgfRe	positive
or (Oriya)	 ଓଡ଼ିଆଙ୍କ ଗଣପର୍ବ #ରଜ, ପ୍ରକୃତି, ନାରୀ ଓ ଧରିତ୍ରୀମାତାଙ୍କ ସୁରକ୍ଷା ସହ କୃତଜ୍ଞତା ପ୍ରଦର୍ଶନ ଲାଗି ଉତ୍ସର୍ଗୀକୃତ। #COVID19 କାରଣରୁ ଏବର୍ଷ "ରଜ ପର୍ବ" ନିଜନିଜ ଘରେ ରହି ନିରାଡମ୍ବର ଭାବେ ପାଳନ କରିବା ସହ ପରିବାର ଆତ୍ମୀୟଙ୍କ ସହ "ପୋଡ ପିଠା"ର ମଜା ନିଅନ୍ତୁ। #ରଜପର୍ବ #RajaParva Happy Raja Parva to all of u https://t.co/N8pE4sEKhw	 The Oriya mass festival is dedicated to showing gratitude for the protection of #Raj, nature, women and earthlings. Thanks to # COVID19, celebrate this year's "Raj Festival" in your own home and enjoy the "Pod Cake" with your family. #Rajparva #RajaParva Happy Raja Parva to all of u https://t.co/N8pE4sEKhw	positive
bg (Bulgarian)	 Починалите от ковид делта в Англия са 6 пъти повече при ваксинираните, в сравнение с тези без ваксина .А сега де ,тоя вирус защо не слуша и не напада НЕВАКСАНИТЕ !Разваля сценария тоя делта щам . https://t.co/5qkfPCjlua	 The number of deaths from covid delta in England is 6 times higher among those vaccinated, compared to those without vaccine. https://t.co/5qkfPCjlua	negative
my (Burmese)	 တွေတဖြုတ်ဖြုတ်သေနေရတဲ့ထဲ ပြေးစရာ လည်း မြေမရှိ လုံခြုံမှုလည်းမရှိ တကယ့်ကို အမှောင်မိုက်ဆုံးနေ့ရက်တွေပဲ" 😔 လို့ အတွေးတွေလဲ ဝင်လာတယ်ဗျာ	 "Our town is really unlucky. People are dying because of Covid. There is no place to run, no land, no security. These are really the darkest days."	negative

TABLE III: Example tweets with most positive or most negative predicted sentiment scores for different languages (translations by Google Translate).

particular one can see the wide variety of languages that is present in publicly available social media data, but that would typically remain unnoticed, because of the lack of suitable natural language processing tools. Our multi-lingual sentiment analysis, on the other hand, provides access to this data without any overhead compared to a mono-lingual system, and, judging from the English translations, does a good job at identifying positive and negative sentiments.

IV. Conclusion

In this work, we studied the task of social media analysis, specifically sentiment analysis, where currently almost exclusively models for English and a very small number of other frequent languages are available. To overcome this obstacle, we propose a method for multi-lingual sentiment classification, which can handle many languages, even rare ones, through the use of multi-lingual sentence representations and through dataset self-annotation.

Our experiments show that the models trained multi-lingually achieve better prediction quality on rare languages than models trained on each language separately, thereby avoiding rare-language discrimination. One remaining shortcoming of our system is that it is not free of biases itself. In particular, the use of Twitter data that was collected based on COVID-19-related keywords, introduces a thematic bias that will presumably make the sentiment predictions less accurate for other forms of data rather than tweets and for texts of unrelated topics, such as product reviews. Note that this is purely because of the datasets used. The proposed method is equally applicable to other textual domains, as long as a self-annotation routine can be constructed.

The use of emoticons for self-annotation also creates a bias as their popularity differs between different age groups and different countries. Furthermore, the specific meaning of emoticons can be context-dependent, which makes it likely that the automatically generated annotation is less precise than that created by human experts.

In future work, we plan to further improve the models, e.g. by using other and more diverse data sources and fine-tuning the underlying representation network. We will then release our code and the trained models for free public use. We also intend to tackle other typical natural language processing tasks besides sentiment analysis in a unified framework.

Acknowledgment

This research was funded in parts by the FORTE program of the Austrian Research Promotion Agency (FFG) and the Federal Ministry of Agriculture, Regions and Tourism (BMLRT) as part of the AMMONIS project (grant no. 879705). The research was also supported by the Scientific Service Units (SSU) of IST Austria through resources provided by Scientific Computing (SciComp).

References

[1] J. Meneghello, N. Thompson, K. Lee, K. W. Wong, and B. Abu-Salih, "Unlocking social media and user generated content as a data source for knowledge management," *International Journal of Knowledge Management (IJKM)*, vol. 16, no. 1, pp. 101–122, 2020.

[2] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A survey of sentiment analysis from social media data," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 450–464, 2020.

[3] F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, *Sentiment analysis in social networks*. Morgan Kaufmann, 2016.

[4] A. Farzindar and D. Inkpen, "Natural language processing for social media," *Synthesis Lectures on Human Language Technologies*, vol. 8, no. 2, pp. 1–166, 2015.

[5] C. W. Leung, "Sentiment analysis of product reviews," in *Encyclopedia of Data Warehousing and Mining, Second Edition*. IGI Global, 2009, pp. 1794–1799.

[6] N. Anstead and B. O'Loughlin, "Social media analysis and public opinion: The 2010 uk general election," *Journal of computer-mediated communication*, vol. 20, no. 2, pp. 204–220, 2015.

[7] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in *International conference on Privacy, Security, Risk and Trust and International Conference on Social Computing*. IEEE, 2011, pp. 192–199.

[8] J. R. Ragini, P. R. Anand, and V. Bhaskar, "Big data analytics for disaster response and recovery through sentiment analysis," *International Journal of Information Management*, vol. 42, pp. 13–24, 2018.

[9] A. Alamoodi, B. Zaidan, A. Zaidan, O. Albahri, K. Mohammed, R. Malik, E. Almahdi, M. Chyad, Z. Tareq, A. Albahri *et al.*, "Sentiment analysis and its applications in fighting covid-19 and infectious diseases: A systematic review," *Expert systems with applications*, p. 114155, 2020.

[10] J. M. Pérez, J. C. Giudici, and F. M. Luque, "pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks," *CoRR*, vol. abs/2106.09462, 2021. [Online]. Available: https://arxiv.org/abs/2106.09462

[11] H. Peng, E. Cambria, and A. Hussain, "A review of sentiment analysis research in chinese language," *Cognitive Computation*, vol. 9, no. 4, pp. 423–435, 2017.

[12] J. Ramteke, S. Shah, D. Godhia, and A. Shaikh, "Election result prediction using twitter sentiment analysis," in *2016 international conference on inventive computation technologies (ICICT)*, vol. 1. IEEE, 2016, pp. 1–5.

[13] M. Coletto, A. Esuli, C. Lucchese, C. I. Muntean, F. M. Nardini, R. Perego, and C. Renso, "Perception of social phenomena through the multidimensional analysis of online social networks," *Online Social Networks and Media*, vol. 1, pp. 14–32, 2017.

[14] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, 2014.

[15] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine learning-based sentiment analysis for Twitter accounts," *Mathematical and Computational Applications*, vol. 23, no. 1, p. 11, 2018.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, Tech. Rep., 2018.

[18] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford*, vol. 1, no. 12, 2009.

[19] G. S. Solakidis, K. N. Vavliakis, and P. A. Mitkas, "Multilingual sentiment analysis using emoticons and keywords," in *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 2. IEEE, 2014, pp. 102–109.

[20] M. Artetxe, G. Labaka, and E. Agirre, "Learning bilingual word embeddings with (almost) no bilingual data," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 451–462.

[21] A. Joulin, P. Bojanowski, T. Mikolov, H. Jégou, and E. Grave, "Loss in translation: Learning bilingual word mapping with a retrieval criterion," *arXiv preprint arXiv:1804.07745*, 2018.

[22] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4512–4525.

[23] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct 2017, pp. 308–317.

[24] D. Dimitrov, E. Baran, P. Fafalios, R. Yu, X. Zhu, M. Zloch, and S. Dietze, "Tweetscov19 - a knowledge base of semantically annotated tweets about the covid-19 pandemic," *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, Oct 2020. [Online]. Available: http://dx.doi.org/10.1145/3340531.3412765

[25] S. A. Memon and K. M. Carley, "Characterizing covid-19 misinformation communities using a novel twitter dataset," 2020.

[26] D. M. W. Powers, "Applications and explanations of Zipf's law," in *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, 1998.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[28] D. Rey and M. Neuhäuser, *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg, 2011, ch. Wilcoxon-Signed-Rank Test, pp. 1658–1659.