Cross-lingual Emotion Intensity Prediction

Irean Navas Alejo
Expert System
irean.navas@gmail.com

Toni Badia
Universitat Pompeu Fabra
tbadia@upf.edu

Jeremy Barnes
University of Oslo
jeremycb@ifi.uio.no

Abstract

Emotion intensity prediction determines the degree or intensity of an emotion that the author expresses in a text, extending previous categorical approaches to emotion detection. While most previous work on this topic has concentrated on English texts, other languages would also benefit from fine-grained emotion classification, preferably without having to recreate the amount of annotated data available in English in each new language. Consequently, we explore cross-lingual transfer approaches for fine-grained emotion detection in Spanish and Catalan tweets. To this end we annotate a test set of Spanish and Catalan tweets using Best-Worst scaling. We compare six cross-lingual approaches, e.g., machine translation and cross-lingual embeddings, which have varying requirements for parallel data – from millions of parallel sentences to completely unsupervised. The results show that on this data, methods with low parallel-data requirements perform surprisingly better than methods that use more parallel data, which we explain through an in-depth error analysis. We make the dataset and the code available at `https://github.com/jerbarnes/fine-grained_cross-lingual_emotion`.

1 Introduction

Emotion analysis within natural language processing attempts to identify the private states (Wiebe et al., 2005) expressed in written text, which in many cases are only implicitly available. Research often classifies these emotions into discrete categories (Ekman, 1999; Plutchik, 2001), such as *anger*, *fear*, *joy*, or *sadness*. This *discrete* approach to emotion has been applied to fairy tales (Alm et al., 2005), headlines (Strapparava and Mihalcea, 2007), and more recently micro-blogging services, such as twitter (Mohammad et al., 2015; Schuff et al., 2017). However, people can express emotion in ways that require a more fine-grained approach than the basic discrete version. Take the following two sentences:

(1) I am not feeling particularly happy today

(2) I feel like I am the most miserable person on earth

Both of these examples would be labelled with the emotion *sadness*. However, it is clear that the second sentence expresses a larger degree of sadness than the first, which categorical approaches to emotion analysis would not be able to identify. This motivates the need to move to a more fine-grained approach to emotion analysis. *Emotion intensity prediction* (Mohammad and Bravo-Marquez, 2017) does just this by extending emotion prediction from a classification task to a regression task. Given a text, the goal is to determine a real-valued number between -1 and 1 representing the *intensity* of the emotion present. This approach allows to capture more subtle differences between expressions of emotion.

Current state-of-the-art approaches to emotion intensity are based on supervised machine learning approaches, which combine several sources of annotated corpora, emotion and sentiment lexicons in order to achieve the best performance. However, the combination of all of these necessary resources is only

Proceedings of the Third Workshop on Computational Modeling of PEople's Opinions, PersonaLity, and Emotions in Social media, pages 140–152
Barcelona, Spain (Online), December 13, 2020.

available in a few high-resource languages, with English easily having the largest number. Collecting a similar set of resources for all other languages is prohibitively expensive and would require years of work. Therefore, it would be preferable to find a way to use the available resources in English to perform emotion intensity prediction in other languages.

Cross-lingual methods – either translation or cross-lingual embedding approaches – offer a possible solution to the lack of labeled data and have shown promise for sentiment analysis at document-level (Chen et al., 2018; Chen et al., 2019), sentence-level (Barnes et al., 2018; Feng and Wan, 2019), and fine-grained (Hangya et al., 2018; Barnes and Klinger, 2019). However, the greater number of classes in emotion classification and the difficulty of the regression task means that it is not obvious that the cross-lingual approaches that work well for sentiment analysis will necessarily work for cross-lingual emotion intensity prediction.

In this work, we provide the first attempt at cross-lingual emotion intensity prediction, by comparing methods which rely on cross-lingual embeddings, machine translation, and unsupervised machine translation to transfer resources from English to predict the emotion in languages that do not have large available datasets or lexicon resources. For testing, we additionally annotate a dataset of tweets in Spanish and Catalan. Our results show that surprisingly unsupervised machine translation is able to outperform both supervised machine translation and cross-lingual embedding methods on these datasets. We additionally perform detailed quantitative and qualitative error analyses of the supervised and unsupervised translation approaches, concluding that while the overall translation quality of the supervised system is better than the unsupervised system, it often does not translate hashtags, which are an important source of emotion information for this task.

In the rest of the paper we discuss related work (Section 2) and provide a description of the datasets (Section 3) and models (Section 4) used for the experiments. We then discuss the results (Section 5) and provide an in-depth analysis of why certain models perform better (Section 6).

2 Related Work

Emotion detection attempts to identify explicitly or implicitly mentioned emotions in a text, either by following a proposed set of basic emotion categories (Plutchik, 1980; Ekman, 1992; Ekman, 1999) or through valence-arousal approaches (Russell, 2003). However, in contrast to other tasks which also attempt to detect evaluative language, such as subjectivity or sentiment analysis, there are relatively few annotated resources, and most of these resources are found only in English (Alm et al., 2005; Aman and Szpakowicz, 2007; Strapparava and Mihalcea, 2007; Schuff et al., 2017). A notable exception is the deISEAR dataset (Troiano et al., 2019), which crowdsources descriptions of emotional events in German.

Annotating categorical emotion is a subjective and complicated task, which often leads to low inter-annotator agreement (Schuff et al., 2017). However, Best-worst scaling has shown to improve overall agreement scores when annotating tweets (Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018). In this approach, annotators are shown n items ($n > 1$, normally 4) and they must choose only the items that are *best* and *worst*, *i. e.* those that most and least represent the phenomena under question. Despite these advances in annotating, for most languages in the world, there exists no annotated emotion dataset which could enable supervised emotion classification.

On English data, previous approaches to classifying emotion have used word and character n-gram features (Mohammad, 2012), sentiment and emotion lexicon features (Mohammad and Kiritchenko, 2015), as well as a variety of neural networks (Köper et al., 2017; Felbo et al., 2017; Bostan and Klinger, 2019). Typically, strong emotion classification systems use a combination of these features to get the strongest performance.

2.1 Emotion intensity prediction

Emotion intensity proposes a more fine-grained view of emotion classification. Specifically, given a tweet and an emotion X, the goal is to determine the intensity or degree of emotion X expressed in the text – a real-valued score between 0 and 1. This task has already been the topic of two shared tasks (Mohammad and Bravo-Marquez, 2017; Mohammad et al., 2018), which attracted many participants.

Given the complexity of the task and the relatively small amount of annotated training data available, it is perhaps unsurprising that state-of-the-art methods incorporate information from external sources, either in the form of specialized word embeddings (Goel et al., 2017), lexicon features (Köper et al., 2017; Duppada and Hiray, 2017), or transfer learning methods (Felbo et al., 2017).

Additionally, in contrast to related tasks, such as sentiment analysis where end-to-end neural methods often give state-of-the-art results (Barnes et al., 2017; Ambartsoumian and Popowich, 2018), for emotion intensity prediction n-grams, character n-grams, word embedding features, and lexicon features play a more important role (Mohammad and Bravo-Marquez, 2017; Köper et al., 2017; Duppada and Hiray, 2017). However, it is not clear if the same features are equally important when performing this task crosslingually.

2.2 Cross-lingual approaches

For other tasks which classify affective text, such as sentiment analysis, cross-lingual approaches have shown promise for classifying a low-resource target language by leveraging labeled data from high-resource source languages, such as English (Barnes et al., 2018; Chen et al., 2019).

We divide cross-lingual approaches into machine translation (MT) techniques and word embedding techniques, as they generally have different data requirements and different models. MT approaches can either be supervised, which requires parallel corpora, or unsupervised, relying only on monolingual corpora) approaches. While most MT research has focused on resource-rich languages where Neural MT (NMT) has indeed displaced Statistical MT, a recent line of work has managed to train a NMT system without any supervision, relying on monolingual corpora alone (Artetxe et al., 2018). This would be particularly useful for low-resource languages if the translation quality proved good enough to enable a classifier to reliable predict the emotion.

Cross-lingual embedding methods instead require large monolingual corpora, and small amounts of bilingual signal, often only small bilingual lexica. Barnes et al. (2018) uses monolingual embeddings, bilingual lexicons and jointly learns cross-lingual embeddings while training an sentiment classifier. The bilingual sentiment embeddings (BLSE) method predicts sentiment of source sentences projecting the vector of the source embeddings into the joint space and repeats the process for the target language using the target embeddings, meaning the original datasets and not the translations, and projecting them into the joint space to obtain the prediction.

More recently, cross-lingual methods have resorted to multilingual language modelling (Devlin et al., 2019; Conneau et al., 2020) based on pretraining large transformer models (Vaswani et al., 2017) on unlabeled text. These models do not explicitly model inter-language representations, but they give surprising cross-lingual performance on many tasks (Wu and Dredze, 2019).

3 Datasets

For training the emotion intensity classifiers, we use the English data from the WASSA 2017 shared task on emotion intensity prediction (Mohammad and Bravo-Marquez, 2017). The authors collected tweets and used crowd-annotation to achieve real-valued labels for four emotions (*anger*, *fear*, *joy*, and *sadness*). We use their predefined splits (statistics are shown in Table 3).

3.1 Annotation of test data

For testing in the two target languages (Spanish and Catalan) we create two annotated datasets following the methodology of Mohammad and Bravo-Marquez (2017). We gather tweets (342 Spanish tweets and 280 Catalan tweets) which contained one of the six emotion terms[1] *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*). This original download leads to between 385-600 tweets per emotion, totalling 3279 Spanish and 2941 Catalan tweets. These tweets are then filtered and normalized in order to delete the retweets, mentions and links, as well as adverts and tweets that only contained images, resulting in 342 tweets in Spanish and 280 in Catalan.

[1] We used the following translations of the emotion terms to gather tweets: *felicidad, enfadado, tristeza, asco, miedo, sorpresa* for Spanish and *felicitat, enfadat, tristesa, fàstic, por, sorpresa* in Catalan.

1.	Que lindo fue volver a meterse a nadar hoy
2.	Hoy estoy triste....
3.	le acabo de romper una pata a la araña y se me subio a la mano
4.	estoy tan enfadado.... gracias a dios que nadie me entiende.

Table 1: An example 4-tuple for the Spanish data for *sadness*.

	Catalan		Spanish	
	Pearson	Spearman	Pearson	Spearman
anger	0.68 (0.02)	0.66 (0.03)	0.77 (0.02)	0.78 (0.02)
disgust	0.71 (0.02)	0.70 (0.02)	0.76 (0.02)	0.77 (0.02)
fear	0.69 (0.02)	0.67 (0.02)	0.74 (0.02)	0.74 (0.02)
joy	0.66 (0.02)	0.64 (0.02)	0.76 (0.02)	0.76 (0.02)
sadness	0.65 (0.02)	0.64 (0.02)	0.74 (0.02)	0.75 (0.02)
surprise	0.44 (0.03)	0.44 (0.04)	0.65 (0.02)	0.62 (0.02)

Table 2: Split-half reliability (as measured by Pearson and Spearman rank correlation) for *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise* annotations of tweets in the Tweet Emotion Intensity Dataset.

Following the methodology set out in Best-Worst Scaling (BWS), each annotator is given four items (4-tuple) and is asked which item is the *best* (highest in terms of the property of interest) and which is the *worst* (least in terms of the property of interest) (Kiritchenko and Mohammad, 2016). The annotator must then choose which one of those 4 tweets represents each emotion the most and which represents it the least. Given the small number of tweets, we include all in the annotation of each emotion. An example of the annotation process in Spanish is shown in Table 1.

This annotation task presents a number of challenges. For example, annotators cannot simply rely on keywords in the tweet to identify the intensity of an emotion, given that many times the authors of tweets use emotional hashtags ironically. Additionally, differentiating between four tweets that all have relatively low intensities of an emotion can be difficult. Finally, inferring the emotion conveyed in a short text is known to be subjective and challenging on its own (Schuff et al., 2017).

3.2 Inter-annotator agreement

After annotating the tuples, we use split-half correlation to determine inter-annotator agreement (shown in Table 2). We report Pearson correlation, which is a number between -1 and 1 that indicates the extent to which two variables are linearly related, and Spearman correlation, which measures the strength and direction of association between two ranked variables. As shown in Table 2, we obtain strong correlations (> 0.6) for all emotions except for *surprise*. We find a higher correlation score in Spanish than in Catalan for all emotions. The lower correlation for the Catalan tweets could be due to the fact that there are fewer tweets referring to *joy* and *surprise*, which made the annotation task harder. It is important to point out that *surprise* has the lowest scores, which is known to be a difficult emotion to classify (Schuff et al., 2017), and has even been split into positive and negative surprise (Alm et al., 2005). In fact, we disregard *surprise* and *disgust* for the rest of the experiments, as we have no English annotated data for these emotions. However, the data will be made available with all annotations.

4 Experimental Setup

In order to determine how much bilingual signal is required to predict crosslingual emotion intensity, we compare four methods with differing data needs. In the following, we describe these methods from those that require the largest amount of bilingual signal (MT) to those that require the least (UNSUP).

	EN$_{train}$	EN$_{dev}$	EN$_{test}$	CA$_{test}$	ES$_{test}$
anger	857	84	760	280	342
fear	1147	110	995	280	342
joy	823	79	714	280	342
sadness	786	74	673	280	342

Table 3: Statistics of the English train (EN$_{train}$), development (EN$_{dev}$), and Catalan (CA$_{test}$) and Spanish (ES$_{test}$) test sets used in the experiments.

Each model is trained on EN$_{train}$ and then tested on EN$_{test}$, CA$_{test}$, and ES$_{test}$. EN$_{dev}$ is only used for hyperparameter optimization.

Supervised Machine Translation (MT): We use GoogleTranslate[2], which makes use of large amounts of parallel data, to translate the test samples to English. We then train a Support Vector Regression model[3] on bag of words representations (MT-BOW) and a second model with a number of additional features (MT-FULL). For the MT-FULL model, we include n-gram features (1-4), character n-grams (3-5), embedding features created by averaging the embeddings of all the tokens in the tweet, and finally features from the following lexicons: NRC Hashtag Sentiment Lexicon (Mohammad et al., 2013), NRC hashtag emotion association lexicon (Mohammad, 2012; Mohammad and Kiritchenko, 2015) and the NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2013), where each feature is a real valued number which represents how much each word is associated to a polarity or emotion. The final representation of each tweet using the MT-FULL method is therefore a 65860 dimensional vector. Finally, we train the SVR model using the following settings (linear kernel, $C = 100$) on the original English training data and test on the translated test set.

Cross-lingual Word Embeddings (CWE): We create 300 dimensional monolingual word2vec embeddings for source and target languages by training on Wikipedia corpora (see UNSUP for more information on the corpora) and then use VecMap (Artetxe et al., 2017) to learn an orthogonal projection of the word embeddings to a joint shared embedding space using a small bilingual lexicon[4] as supervision (5749 and 5310 translation pairs for EN-ES and EN-CA, respectively). Finally, we train a Support Vector Regression model on the source language (EN) using only the crosslingual embeddings as features using the following settings (linear kernel, $C = 100$) and test it on the target languages (ES, CA). It is important to highlight that this method does not use the translations but the original texts.

Bilingual Sentiment Embeddings (BLSE): Like CWE, this model uses a bilingual lexicon to learn a mapping from both original vector spaces to a shared bilingual space, but instead jointly learns to predict the sentiment and employs two linear projection matrices. This allows the model to infuse the target embedding space with sentiment information by updating the source space for sentiment and requiring that the target space resemble it as much as possible, using the bilingual dictionary to anchor terms. In this work, we adapt the model to predict emotion intensity by replacing the cross-entropy loss with mean-squared error. We train the model on the English training data and the same bilingual lexicons as for CWE, optimizing with Adam (Kingma and Ba, 2014) for 100 epochs with an α of 0.001. We keep the model with the best performance on the source language development set, and finally test on the target test set. As with CWE, we highlight that this method does not use the translations but the original texts.

Unsupervised Statistical Machine Translation (UNSUP) We train an unsupervised statistical machine translation model (Artetxe et al., 2018) on Wikipedia corpora for both English-Spanish and

[2] Available at `https://translate.google.com`.

[3] We use the version available in the Sklearn toolkit (Pedregosa et al., 2011).

[4] We use the lexicons made available from Barnes et al. (2018).

	Monolingual					Cross-lingual									
	English					Catalan					Spanish				
	anger	*fear*	*joy*	*sadness*	avg.	*anger*	*fear*	*joy*	*sadness*	avg.	*anger*	*fear*	*joy*	*sadness*	avg.
Cwe	0.17	0.30	0.22	0.28	0.24	0.04	-0.02	0.13	0.03	0.05	0.03	-0.04	0.16	0.00	0.04
Blse	0.35	0.27	0.46	0.39	0.37	0.24	-0.06	0.03	0.06	0.07	0.14	0.09	0.24	**0.12**	0.15
Mt-Bow	0.41	0.51	0.48	0.47	0.47	0.28	0.10	0.19	0.02	0.15	0.14	0.06	0.17	-0.11	0.07
Unsup-Bow	0.41	0.51	0.48	0.47	0.47	0.21	-0.03	0.17	0.03	0.10	0.14	0.01	0.13	-0.10	0.05
Mt-Full	**0.60**	0.60	**0.64**	**0.62**	**0.62**	0.37	0.35	**0.46**	0.17	0.34	0.33	0.24	0.43	0.05	0.26
Unsup-Full	**0.60**	0.60	**0.64**	**0.62**	**0.62**	**0.42**	0.40	0.45	0.21	**0.37**	**0.42**	**0.31**	**0.43**	0.10	**0.32**
mBert	0.20	0.27	0.33	0.31	0.27	-0.05	0.06	0.03	-0.15	-0.03	0.12	0.13	0.05	0.0	0.08
XLM-Roberta	0.44	**0.61**	0.54	0.58	0.54	0.35	**0.42**	0.39	**0.25**	0.35	0.34	0.29	**0.43**	**0.12**	0.30

Table 4: Pearson results of monolingual English-English experiments, as well as cross-lingual English-Catalan and English-Spanish for each emotion and each model. Average column added and best results are shown in **bold**.

English-Catalan. The model first creates monolingual embeddings, then learns to project them to a bilingual space by selecting identical strings as pivots, which serve as a noisy bilingual lexicon, which is improved iteratively. Next, the model induces a noisy phrase table for the SMT model, which is again improved iteratively.

We extract cleaned corpora[5] from Wikipedia dumps and sentence and word tokenize them, resulting in 89~/29~/10~ million sentences for English, Spanish, and Catalan, respectively. We train the Unsup model using the default settings (removing sentences with fewer than 3 and more than 80 tokens, 5-gram language model, 300 dimensional embeddings, 10 rounds of unsupervised tuning for the SMT and 3 rounds of backtranslation). We then translate the test data to English using the Unsup system. Finally, we use bag-of-words representations (Unsup-Bow) and additional n-gram, character n-gram, embedding, and lexicon information (Unsup-Full) and train a Support Vector Regression model using the following settings (linear kernel, $C = 100$), as we do with the Mt models.

Multi-lingual Language Models We use pretrained mBert and XLM-Roberta models to extract features for each example by taking the final `[CLS]` embedding as the representation for the example. These features are then used to train an SVR model, as with the other experiments. We additionally experimented with adding a linear layer after the final LM layer and fine-tuning the full model, only fine-tuning the linear layer, and using a max pooled representation instead of the `[CLS]` embedding, but found that these approaches did not perform as well.

5 Results

The Pearson correlation results are summarized for all models in Table 4 for Spanish and Catalan. We report the individual scores for *anger*, *fear*, *joy*, and *sadness*, as well as the averaged Pearson score of all emotions.

The approach that obtains the highest overall Pearson correlation across all emotions on both languages is Unsup-Full, averaging 0.37 on Catalan and 0.32 on Spanish. In addition, it is the best performing model on 4 of the 8 tasks, except for Catalan *joy*, where Mt-Full is 0.01 percentage points (pp.) better, reaching 0.46 and Catalan *sadness* (XLM-Roberta is 0.04 pp. better, at 0.25) and Catalan *fear* (XLM-Roberta is 0.02 pp. better, at 0.42) and Spanish *sadness* (Blse is 0.02 pp. better, reaching 0.12). XLM-Roberta is the second best model, averaging 0.35 and 0.30 on Catalan and Spanish, respectively, while Mt-Full is slightly worse (0.34 and 0.26). Mt-Bow and Unsup-Bow perform much worse (0.15/0.07 and 0.10/0.05), and the cross-lingual embedding methods are the worst by far (0.04/0.03 for Cwe and 0.07/0.15 for Blse).

[5] We use the wikiextractor tool available at `https://github.com/attardi/wikiextractor`.

		hashtags	lexical	insert.	delet.	untrans.	slang	names	nums.	Total
CA	MT	**90**	53	2	18	17	26	5	2	213
	UNSUP	60	**67**	**7**	**14**	**168**	29	**81**	**9**	**435**
ES	MT	**62**	37	0	4	12	68	0	0	183
	UNSUP	35	**142**	**13**	**43**	**84**	**101**	**49**	**16**	**467**

Table 5: Error analysis of the the machine translation used in MT and UNSUP approaches. The error categories include incorrectly translated hashtags, lexical errors, insertions, deletions, untranslated segments, translation errors of slang and non-standard language, mistranslated names and numbers. Number refer to the number of tweets where these errors are found, rather than the number of errors.

It is clear that the additional features (character n-grams, embedding features, and lexicon features) are essential. MT-FULL performs an average of 0.19 pp. Pearson better than MT-BOW, while UNSUP-FULL leads to 0.28 pp. improvement over UNSUP-BOW. We further confirm this in Section 6.2.

Regarding the cross-lingual embedding models, it seems evident that these do not contain enough information to accurately predict emotion intensity in the target language. BLSE does outperform CWE on both Catalan and Spanish (an average of 0.03 pp. and 0.12 pp., respectively) and both MT-BOW and UNSUP-BOW on Catalan (0.08 pp. and 0.10 pp.), but the overall performance is still poor. These models are also the poorest performers monolingually.

There is large performance gap between XLM-ROBERTA and MBERT, on both the monolingual (0.27 pp.) and cross-lingual tasks (0.38/0.22 pp.), as MBERT is the weakest cross-lingual model and XLM-ROBERTA the second best.

Additionally, there is a divergence between *sadness* and the rest of the emotions analyzed, with no model achieving more than 0.21 or 0.12 in Catalan and Spanish. This seems to indicate that sadness may be harder to classify cross-lingually, as monolingually this class is has the best classification results (Mohammad and Bravo-Marquez, 2017; Köper et al., 2017). This class also has the fewest training and development examples in English, which may indicate that the good previous results monolingually may have been due to overfitting to the data. It is also possible that the particulars of the target language test data are the reason for this difference, although the inter-annotator agreement scores suggest that *sadness* is not more difficult than the other classes.

6 Analysis

In this section we compare both quantitatively and qualitatively the differences in translation quality between MT and UNSUP. Furthermore, we perform an ablation study to determine which features are the most important for MONO, MT-FULL, and UNSUP-FULL.

6.1 Differences in translation quality

Given that twitter is a social network where people express their emotions and opinions on a large variety of topics – social or personal events, news, and politics – the translation task is made more difficult. Additionally, relevant information to emotion classification in tweets is often contained in hashtags (Mohammad et al., 2013), which are known to be difficult to translate (Gotti et al., 2014). Therefore, cross-lingual approaches to fine-grained emotion detection in twitter are particularly challenging since the language used in twitter usually contains abbreviations, acronyms, emoticons, unusual orthographic elements, slang, and misspellings (Liew and Turtle, 2016). All of these phenomena are difficult for both translation- and projection-based cross-lingual approaches.

We manually examine the MT and UNSUP translations of the Catalan and Spanish tweets for translation errors. For each tweet, we determine if there has been an error regarding the hashtags, any lexical errors, insertions, deletions, untranslated segments, errors with non-standard language, errors translating

original	*#DiosLosCríaYEllosSeJuntan* L'advocat de Camacho en el cas Método 3 va redactar la sentència de Puig Antich link
MT	*# DiosLosCríaYesLocated* The Camacho lawyer in the case Method 3 wrote the sentence of Puig Antich link
UNSUP	*# DiosLosCríaYEllosSeJuntan* l 'advocat of camacho in the case história 3 drafted the verdict of abu-jamal link
manual trans.	*#BirdsOfAFeatherFlockTogether* Camacho's lawyer in the Método 3 case is the one who sentenced Puig Antich link

Table 6: An example of a tweet in Catalan (original), its translations using the two machine translation systems (MT, UNSUP), as well as a manual translation. *Untranslated tokens* are highlighted in red, while *entity errors* are highlighted in blue.

original	Harto de la situación en #*Cataluña*. Votamos mayoritariamente a delincuentes y tenemos lo que merecemos. No hay solución #*verguenza*
MT	Fed up of the situation in # *Catalonia* . We vote mostly criminals and we have what we deserve. There is no solution *# verguenzak*
UNSUP	Fed up with the situation in # *catalonia* . They voted overwhelmingly to criminals and we have what they deserve no solution # disgrace
manual trans.	Tired of the situation in #*Catalonia*. We mainly vote for criminals and get what we deserve. There's no solution. #*shame*

Table 7: An example of tweet in Spanish (original) and its translation using the two machine translations systems (MT, UNSUP). *Hashtag translation errors* are highlighted in grey and *lexical errors* are highlighted in green.

names and errors translating number and show the results in Table 5. MT has fewer errors overall compared to UNSUP (213/183 compared to 435/467, respectively) and has fewer of all error types, except for hashtags. For the task of predicting emotion intensity in tweets, the hashtags are often the most informative source, which explains why UNSUP-FULL performs better than MT-FULL in our experiments.

The Spanish translation models generally perform better than the Catalan ones. This is likely due to the larger amount of training data available. However, the Spanish data also contains more use of non-standard language, which is reflected in the *slang* errors. In these cases, MT generally performs much better than UNSUP. Interestingly, UNSUP tends to mistranslate named entities. Specifically, the model often replaces a named entity with a *similar* entity in the target language. For example, mentions of the Catalan freedom fighter Salvador Puig i Antich are consistently translated to Mumia Abu-Jamal, an American journalist (see Table 6). Both were accused of killing a police officer and sentenced to death, which lead to large protests. This is likely due to the nearest neighbor search used to create the original phrase tables.

Besides the mistranlation of named entities, in Table 6 we can also see that the multiword hashtag, which contains information necessary to properly interpret the emotional content of the tweet, has not been translated by MT or UNSUP. Note that although this problem could be improved by properly segmenting the hashtags in a previous step (Declerck and Lendvai, 2015; Çelebi and Özgür, 2016), translation would still likely lead to a loss of information (Gotti et al., 2014) important for emotion classification.

Table 7, instead, shows an example from the Spanish dataset where, even though the MT version better preserves the semantics of the original tweet, it did not correctly translate the emotional hashtag, while UNSUP did. On the other hand, for cases where MT-FULL has better performance, translation quality tends to be the main factor. Specifically, UNSUP-FULL tends to leave many words untranslated.

		ALL	-ngrams	-char	-embs	-hashtag	-emo	-sent	-all lex
anger	Mono	0.60	**-0.31**	-0.04	-0.01	-0.06	-0.01	-0.01	-0.06
	Mt-Full	0.37	**-0.30**	-0.00	-0.00	-0.02	-0.00	-0.02	-0.07
	Unsup-Full	0.42	**-0.23**	-0.05	-0.00	-0.02	-0.01	-0.05	-0.11
fear	Mono	0.60	**-0.30**	-0.01	-0.00	-0.01	-0.00	-0.01	-0.04
	Mt-Full	0.35	**-0.31**	+0.06	-0.00	-0.02	-0.01	-0.05	-0.14
	Unsup-Full	0.40	**-0.26**	+0.03	-0.00	-0.01	-0.01	-0.08	-0.21
joy	Mono	0.64	-0.03	-0.03	-0.00	-0.00	-0.02	-0.03	**-0.06**
	Mt-Full	0.46	-0.07	-0.02	-0.00	-0.01	-0.02	-0.02	**-0.11**
	Unsup-Full	0.45	-0.05	-0.03	-0.00	-0.01	-0.02	-0.04	**-0.16**
sadness	Mono	0.62	-0.01	-0.02	-0.00	-0.00	-0.00	-0.01	**-0.04**
	Mt-Full	0.17	-0.01	-0.01	-0.00	-0.01	-0.00	-0.09	**-0.15**
	Unsup-Full	0.21	-0.02	+0.05	-0.00	-0.02	-0.02	-0.06	**-0.20**

Table 8: Ablation study of Mono (used to show an informative monolingual baseline), Mt-Full and Unsup-Full on the Catalan dataset, where we show the drop in performance (Pearson correlation) when we remove only a single feature at a time (except for -all lex, where all lexicon features are removed). We show the largest drop in **bold** and the second largest underlined. On most emotions (*anger*, *fear joy*), removing the n-gram feature leads to the largest drop both mono- and cross-lingually. For *sadness*, however, the NRC sentiment lexicon features (sent) are most decisive.

6.2 Ablation study

In order to determine which features are most predictive for emotion intensity, we perform an ablation study of Mono, Mt-Full, and Unsup-Full on the Catalan test data[6]. Specifically, we remove a single feature at at time, except for -all lex, where all lexicon features are removed. We include Mono as an upper-bound to determine what features are most important for the task, given enough monolingual data, but note that the test data is different for Mt-Full and Unsup-Full, so the exact results are therefore not strictly comparable. The results are shown in Table 8.

In general, the cross-lingual models exhibit the same relationship to the features as the monolingual model, although with generally lower performance. Token n-grams are the most important feature for *anger* and *fear*, although less important for *joy* and *sadness*. Word embedding features seem to contribute nothing to the performance. Character n-gram features, on the other hand, contribute little, or even hurt the performance (removing them actually improves the results for Mt-Full and Unsup-Full on *fear*, and for Mono and Unsup-Full on *sadness*). Finally, the lexicon features are important for all emotions.

For *joy*, removing any one set of features does not lead to large drops in performance. Given the good performance of all models on this emotion, it seems to indicate that this class is the easiest to predict, and that the features are relatively redundant. However for Unsup-Full, removing all lexicon features still leads to a drop of 0.16 pp.

Sadness is the most difficult emotion, with the cross-lingual models performing on par with Mono. The lexicon features are the most informative for all models, specifically the NRC sentiment lexicon features (-sent). This effect is even stronger cross-lingually, where without them, the models perform at chance level.

[6] The ablation study results for the Spanish data are similar.

7 Conclusion

In this paper, we provided the first attempt at cross-lingual emotion intensity prediction, by comparing methods which rely on differing amounts of cross-lingual signal, ranging from millions of parallel sentences (MT), small bilingual dictionaries (cross-lingual embeddings), to no explicit cross-lingual signal at all (UNSUP). We compare these methods on two target languages, Spanish and Catalan, which do not have large available emotion datasets or lexicon resources. In order to test the models, we additionally annotated a small dataset of tweets in Spanish and Catalan.

Our results show that translation methods outperform embedding-based methods for almost all emotions and achieve reasonable average results, although there is still a noticeable gap to reach monolingual levels. Surprisingly, unsupervised translation is the best performing cross-lingual method, largely due to the fact that it more accurately translates hashtags. XLM-ROBERTA performs nearly as well, but unfortunately cannot be combined with sentiment and emotion lexicons available in English, as it processes the original target-language data. These results may not hold for other domains, such as literature or opinion pieces, where emotional information is not concentrated in a similar way.

In the future, it would be interesting to perform experiments on various domains, in order to determine whether unsupervised machine translation for cross-lingual emotion is robust to domain shift. Theoretically, this is simpler for unsupervised MT rather than supervised MT, which could motivate further research in this direction. As lexicon information has proven so useful for this task, it could be interesting to look into approaches that use this information to improve pretrained multilingual language models. Additionally, given the importance of hashtags for emotion detection in tweets, it would be important for future work in cross-lingual emotion detection to concentrate on achieving better translations of hashtags.

Finally, we contemplate promising research on emotion detection and classification using the newly annotated data in Catalan and Spanish we introduce here. We expect that this will contribute to furthering research on these two languages.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, BC, Canada, October.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue: 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007. Proceedings*, pages 196–205. Springer.

Artaches Ambartsoumian and Fred Popowich. 2018. Self-attention: A better building block for sentiment analysis neural network classifiers. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 130–139, Brussels, Belgium, October.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. 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)*, pages 451–462, Vancouver, Canada, July. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium, October-November. Association for Computational Linguistics.

Jeremy Barnes and Roman Klinger. 2019. Embedding projection for targeted cross-lingual sentiment: Model comparisons and a real-world study. *Journal of Artificial Intelligence Research*, 66:691–742.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2017. Assessing State-of-the-Art Sentiment Models on State-of-the-Art Sentiment Datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–12, Copenhagen, Denmark, September.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2483–2493, Melbourne, Australia, July. Association for Computational Linguistics.

Laura Ana Maria Bostan and Roman Klinger. 2019. Exploring fine-tuned embeddings that model intensifiers for emotion analysis. In *Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 25–34, Minneapolis, USA, June. Association for Computational Linguistics.

Arda Çelebi and Arzucan Özgür. 2016. Segmenting hashtags using automatically created training data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2981–2985, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy, July. 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, July. Association for Computational Linguistics.

Thierry Declerck and Piroska Lendvai. 2015. Processing and normalizing hashtags. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 104–109, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

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, June. Association for Computational Linguistics.

Venkatesh Duppada and Sushant Hiray. 2017. Seernet at EmoInt-2017: Tweet emotion intensity estimator. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 205–211, Copenhagen, Denmark, September. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, pages 169–200.

Paul Ekman. 1999. Basic emotions. In Tim Dalgleish and M. J. Powers, editors, *Handbook of Cognition and Emotion*. Wiley.

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, September. Association for Computational Linguistics.

Yanlin Feng and Xiaojun Wan. 2019. Learning bilingual sentiment-specific word embeddings without cross-lingual supervision. 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 420–429, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Pranav Goel, Devang Kulshreshtha, Prayas Jain, and Kaushal Kumar Shukla. 2017. Prayas at EmoInt 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 58–65, Copenhagen, Denmark, September. Association for Computational Linguistics.

Fabrizio Gotti, Phillippe Langlais, and Atefeh Farzindar. 2014. Hashtag occurrences, layout and translation: A corpus-driven analysis of tweets published by the Canadian government. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2254–2261, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Viktor Hangya, Fabienne Braune, Alexander Fraser, and Hinrich Schütze. 2018. Two methods for domain adaptation of bilingual tasks: Delightfully simple and broadly applicable. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 810–820, Melbourne, Australia, July. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, dec.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817, San Diego, California, June. Association for Computational Linguistics.

Maximilian Köper, Evgeny Kim, and Roman Klinger. 2017. IMS at EmoInt-2017: Emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57, Copenhagen, Denmark, September.

Jasy Suet Yan Liew and Howard R. Turtle. 2016. Exploring fine-grained emotion detection in tweets. In *Proceedings of the NAACL Student Research Workshop*, pages 73–80, San Diego, California, June. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark, September.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing Management*, 51(4):480 – 499.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June. Association for Computational Linguistics.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, November.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1(3):3–33.

Robert Plutchik. 2001. The nature of emotions. *American Scientist*, 89(July-August):344–350.

James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):1–145.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark, September. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June.

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. Crowdsourcing and validating event-focused emotion corpora for German and English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, Florence, Italy, July. Association for Computational Linguistics.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China, November. Association for Computational Linguistics.