

# UniSent: Universal Adaptable Sentiment Lexica for 1000+ Languages

Ehsaneddin Asgari<sup>†,‡,◊</sup>, Fabienne Braune<sup>†</sup>, Benjamin Roth<sup>◊</sup>, Christoph Ringlstetter<sup>‡</sup> and Mohammad R.K. Mofrad<sup>†,◊</sup>

<sup>†</sup>Department of Bioengineering, University of California, Berkeley, CA 94720, USA

<sup>‡</sup>NLP Expert Center, Volkswagen Group Data:Lab, Munich, Germany

<sup>◊</sup>Center for Information and Language Processing, Munich 80538, Germany

<sup>◊</sup>Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Lab, Berkeley, CA 94720, USA

## Abstract

In this paper, we introduce *UniSent* universal sentiment lexica for 1000+ languages. Sentiment lexica are vital for sentiment analysis in absence of document-level annotations, a very common scenario for low-resource languages. To the best of our knowledge, *UniSent* is the largest sentiment resource to date in terms of the number of covered languages, including many low resource ones. In this work, we use a massively parallel Bible corpus to project sentiment information from English to other languages for sentiment analysis on Twitter data. We introduce a method called *DomDrift* to mitigate the huge domain mismatch between Bible and Twitter by a confidence weighting scheme that uses domain-specific embeddings to compare the nearest neighbors for a candidate sentiment word in the source (Bible) and target (Twitter) domain. We evaluate the quality of *UniSent* in a subset of languages for which manually created ground truth was available, Macedonian, Czech, German, Spanish, and French. We show that the quality of *UniSent* is comparable to manually created sentiment resources when it is used as the sentiment seed for the task of word sentiment prediction on top of embedding representations. In addition, we show that emoticon sentiments could be reliably predicted in the Twitter domain using only *UniSent* and monolingual embeddings in German, Spanish, French, and Italian. With the publication of this paper, we release the *UniSent* sentiment lexica.

## 1 Introduction

Language technologies permeate our everyday life through web search, translation, online shopping, email writing, spell checking systems etc. The existence of such technologies highly depends on the existence of the underlying computational linguistic resources for a language. Computational linguistic resources such as machine-

readable lexica, part-of-speech-taggers and dependency parsers are available for at most a few hundred languages. This means that the majority of the 7000 languages of the world are low-resource. This gap between advances in language technologies for English versus other languages endangers multilingualism in the digital age. Languages with lack of technological support (as a result of having limited resources) are less used over time and eventually get in danger of extinction. Many EU and US programs are designed to address this issue (Cieri et al., 2016). The rationale of these projects is that even “small” languages are important for the preservation of the common heritage of humankind and cultural diversity which benefits everybody. In addition, certain low-resource languages can be also politically and economically important.

Large amount of text available on the web and applications in marketing (Bollen et al., 2011), social science (Hopkins and King, 2010), political science (Wang et al., 2012; Wong et al., 2016) motivates sentiment analysis of news, blogs, social networks, reviews, opinions, and recommendations. However, sentiment analysis requires either word or document level sentiment annotations. Typically, these are available only for a limited number of languages, preventing accurate sentiment classification in low resource setups. In these scenarios sentiment lexica are important game changers because in many cases end-to-end sentiment classification is not feasible due to a lack of document level annotations (Jurafsky and Martin, 2014).

Recently, embedding-based approaches for supervised or semi-supervised word sentiment inference became popular allowing for lexicon vocabulary expansion and implicit domain adaptation (Rothe et al., 2016; Hamilton et al., 2016). Although using the embedding space as represen-

tation in word sentiment classification sufficiently addresses domain adaptation in many cases, it can be improved in situations where the domain shift for some words from the lexicon seeds in the source vocabulary (e.g. Bible<sup>1</sup>) to the target vocabulary (e.g. Twitter) is large. For instance, in Biblical texts, the Spanish word *sensual* has the connotation of *sin* which has a negative polarity. But in the Twitter domain, the same word is associated with *sexy*, which has a positive polarity. In cases where the classifier using the embedding space fails to capture this shift, enhancing the model via mitigation of the domain mismatches is required.

**Contributions:** We release the first sentiment lexicon covering 1000+ languages and achieving macro-F1 over 0.75 on word sentiment prediction for most evaluated languages, meaning that we enable sentiment analysis in many low resource languages. The creation of *UniSent* requires only a sentiment lexicon in one language (e.g. English) and a small, but massively parallel corpus in a specific domain. We evaluate *UniSent* for word sentiment classification of Macedonian, Czech, German, Spanish, French against manually assigned sentiment polarities and show that its quality is comparable to the use of manually created resources, which is a great evidence that *UniSent* works well also for low-resource languages where we do not have resources for evaluation. Secondly, we evaluated *UniSent* w.r.t. the classification of emoticon sentiments in the Twitter domain, where macro-F1 of 0.79, 0.76, 0.74, and 0.76 were obtained for German, Italian, French, and Spanish respectively.

To ensure the usability of our lexica for any new domain, we propose *DomDrift*, a method requiring only a pretrained embedding space in the target domain, which is relatively a cheap resource to obtain. By comparing the source and target embedding graphs *DomDrift* quantifies the semantic changes of words in the sentiment lexicon in the new domain. This measure can hence be used to weight words in the sentiment lexicon for downstream supervised or semi-supervised sentiment analysis models. We show that on top of implicit domain adaptation, using target domain embeddings, the incorporation of domain drift scores improves sentiment classification for French, Span-

<sup>1</sup>Massively parallel corpora mainly exist in the Bible domain and for a smaller text size for the universal declaration of human rights (Emerson et al., 2014).

ish, and Macedonian.

## Related Work

Several research efforts tackled the automatic creation of sentiment lexica for a multitude of languages, but these efforts resulted in the creation of resources for at most 136 languages (Chen and Skiena, 2014) or in lexicon covering a very specific low-resource language (Afli et al., 2017; Darwich et al., 2017). Moreover, these approaches heavily rely on linguistic resources, such as WordNet or fully trained machine translation systems, which limit them to the languages where these are available. An alternative to our approach in lexicon creation for sentiment is using minimal bilingual supervision (Hangya et al., 2018; Barnes et al., 2018a,b) to create document-level annotations for end-to-end sentiment classifications of documents. Later approaches only work in an end-to-end fashion and do not allow to directly create sentiment lexica.

The annotation projection to create *UniSent* sentiment lexica is inspired by *SuperPivot* introduced in (Asgari and Schütze, 2017) for the typological analysis of tense in 1000 languages. Agić et al. (2016) also use massively parallel corpora to project POS tags and dependency relations across languages. In contrast to these studies, here we perform parallel projection on sentiment information for resource creation and not for typological analysis. In addition, we propose a method called *DomDrift* to mitigate the huge domain mismatch between Bible and target domain via an embedding-based confidence weighting scheme.

## 2 Methods

In the next sections, we describe (i) the main resources required for *UniSent* and (ii) the steps of its creation and adaptation to new domains. The overview of these steps is also depicted in Figure 1.

### 2.1 UniSent Required Resources

**Super-parallel corpus:** The dataset we will work with is the Parallel Bible Corpus (PBC). PBC consists of translations of the New Testament in 1242 languages covering an order of magnitude more languages than any other parallel corpus currently in use in natural language processing research (Mayer and Cysouw, 2014).

**Initial sentiment seeds:** We use a high-quality English sentiment lexicon called WKWSCI (Khoo and Johnkhan, 2018) as a resource to be projected on other languages.

## 2.2 UniSent and DomDrift

Our contributions are two-fold (i) creation of *UniSent* using a cross-lingual projection of sentiment polarities, which needs to be done only once (ii) introducing *DomDrift* a novel method for adapting *UniSent* to any newly observed domain by measuring the domain-drift of words in the new domain. The first part needs a sentiment lexicon in one language (here WKWSCI for English) as well as a massively parallel corpus (PBC). For the second part, only a pre-trained embedding space in the target domain is required. In the next sections, we illustrate our method by creating *UniSent* for one example language (for better readability). The described steps are however repeated for each of the 1000+ languages composing *UniSent*. The steps are detailed next and illustrated in Figure 1.

**(i) UniSent creation using cross-lingual projection of sentiment polarities:** We project sentiment polarities from English (source language) to a target language in the parallel corpus using *SuperPivot* (Asgari and Schütze, 2017). This method projects annotations across 1000+ languages via an alignment graph generated using FastAlign (Dyer et al., 2013) on the PBC corpus. In the FastAlign word alignment pairs ( $w_{\text{source}}, w_{\text{target}}$ ), we replace the source words with their sentiment labels from WKWSCI (where available). Subsequently, we search for the words in the target language that are highly correlated with each of the sentiment labels (positive or negative). We use FDR corrected two-sided  $\chi^2$  (Casella and Berger, 2002) score to find these sentiment seeds. We denote the vocabulary of the target language in the parallel corpus by  $V_{l_t, d_s}$ , where  $l_t$  is the target language and  $d_s$  the source domain (here the domain of the parallel corpus, i.e. biblical domain).<sup>2</sup> This first step generates, in each target language, pairs  $(w_{l_t, d_s}, y)$ , where  $w_{l_t, d_s} \in V_{l_t, d_s}$  is a word in the target language and source domain and  $y$  is a highly correlated sentiment annotation with  $w_{l_t, d_s}$ . Vocabulary  $V_{l_t, d_s}$  is limited to the words in the par-

<sup>2</sup>We call this domain *source domain* because it will be adapted to a target domain in the subsequent steps.

allel corpus<sup>3</sup>. Because most existing super-parallel corpora are from the Bible (Mayer and Johannsen, 2016),  $V_{l_t, d_s}$  besides of being limited in size has the major drawback of originating from a very specific domain. To overcome this limitation, we define a method to measure domain drifts. In addition, we leverage word embeddings as used in (Rothe et al., 2016) to propagate the annotations  $y$  to words of a larger vocabulary than  $V_{l_t, d_s}$ . The overview of *UniSent* creation is depicted in Figure 1.a. The *UniSent* lexica and a complete list of 1242 unique languages<sup>4</sup> covered by *UniSent* along with their language family information are provided in the supplementary material.

**(ii) DomDrift: unsupervised measurement of domain drift:** Word embeddings trained with an unsupervised language modeling objective (e.g., skip-gram) are known to preserve the syntactic and the semantic word similarities in the embedding space (Mikolov et al., 2013; Pennington et al., 2014). Domain changes will certainly impact the neighborhoods in the embedding space. Thus a comparison of words relative distances in two embedding spaces can be used to measure their degree of domain shift (Kulkarni et al., 2015; Asgari and Mofrad, 2016). The main purpose of *DomDrift* is to identify words in the sentiment lexicon having a domain shift in the target domain by comparison of their neighbors in the embedding spaces.

*DomDrift* quantifies the domain drift of a given word in the sentiment lexicon regardless of its label, only by comparison of neighbors in the source and target domains' embedding spaces  $\Omega_{l_t, d_s} : V_{l_t, d_s} \rightarrow \mathbb{R}^{h_s}$ , and  $\Omega_{l_t, d_t} : V_{l_t, d_t} \rightarrow \mathbb{R}^{h_t}$ , where  $h_s, h_t$  are the sizes of the source and target embedding spaces. *DomDrift* quantifies word domain drift as follows:

- (i) In each embedding space, we compute, for each word  $w_{l_t, d_x}$  in the *UniSent* lexicon, the distance distribution  $P(w_{l_t, d_x}, \mathbf{w}_{\Omega_{s,t}})$  of word  $w_{l_t, d_x}$  with all other words in intersection of source and target embedding spaces, i.e.  $\forall w_j \in \mathbf{w}_{\Omega_{s,t}}$ . For this, we take the  $l_1$  normalized cosine distance of the representation of  $w_{l_t, d_x}$  with all other words in the embedding space. This distribution can

<sup>3</sup>Because we use this corpus for the cross-lingual projection

<sup>4</sup>We consider two languages different if they have different ISO 639-3 codes

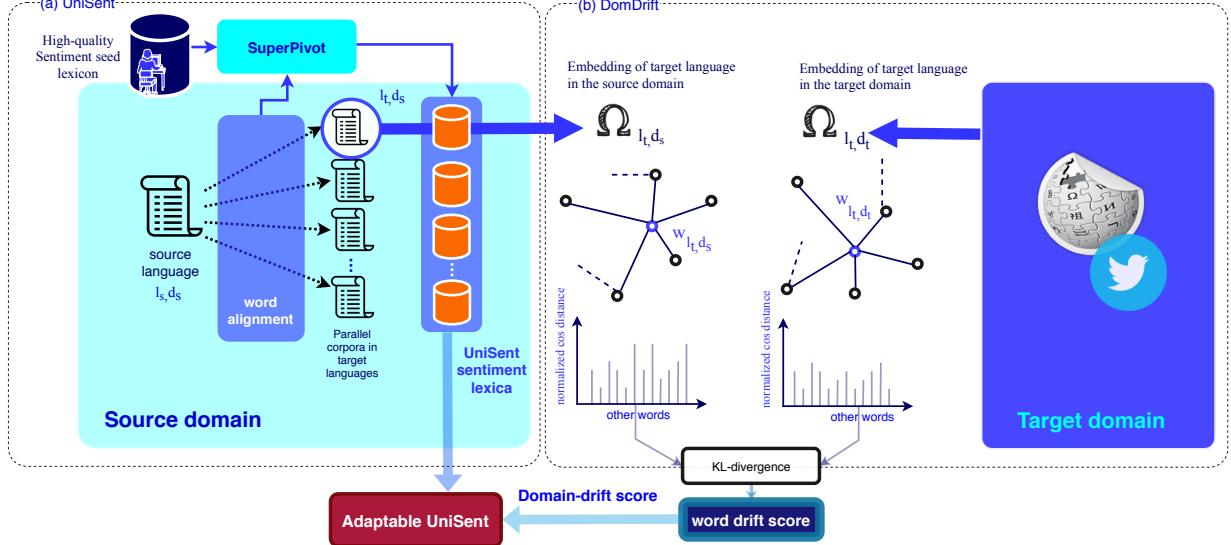


Figure 1: The overview of universal adaptable sentiment lexica. The approach can be divided into two main steps: (a) UniSent creation using SuperPivot method, (b) DomDrift for measurement of word domain drifts.

be regarded as word profile in the domain  $x$  (source or target):

$$P_i(w_{l_t, d_x}, \mathbf{w}_{\Omega_{s,t}}) = \frac{1 - \cos(\vec{w_{l_t, d_x}}, \vec{w_i})}{\sum_j [1 - \cos(\vec{w_{l_t, d_x}}, \vec{w_j})]},$$

where  $w_i, w_j \in \Omega_{s,t}$  and  $\vec{w_k}$  is the embedding representation of word  $w_k \in \Omega_{s,t}$  in domain  $x$  (source or target). We use  $\mathbf{w}_{\Omega_{s,t}}$  so that the word profiles in the source and target domains are comparable, i.e. they have the same elements.

- (ii) We compute, for each word  $w_{l_t, d_x}$  a shift weight between vocabularies  $V_{l_t, d_s}$  and  $V_{l_t, d_t}$  of source and target spaces. This is done by comparing, for each word in the UniSent lexica, its profile in the source  $P(w_{l_t, d_s}, \mathbf{w}_{\Omega_{s,t}})$  and target domain  $P(w_{l_t, d_t}, \mathbf{w}_{\Omega_{s,t}})$  using the Kullback-Leibler divergence.

More formally for a word  $w'$  its domain drift ( $\lambda_{w'}$ ) between the source and the target domains can be calculated as follows.

$$\lambda_{w'} = D_{KL}(P(w'_{l_t, d_s}, \mathbf{w}_{\Omega_{s,t}}) \| P(w'_{l_t, d_t}, \mathbf{w}_{\Omega_{s,t}}))$$

The steps of DomDrift are illustrated in Figure 1.b. As also depicted in the figure, the calculated weights enhance the universal sentiment lexica resulting in a final adaptable version (Adaptable UniSent), e.g. the weights will be used to reduce the influence of huge domain

mismatches in a confidence weighting scheme. In Figure 2 we illustrate an example of domain drift and explain the workings of our weighting method in §3.4. Once we computed our shift weights, they can be used in any semi-supervised or supervised approach (e.g., sample weights in the logistic regression model).

**Source embedding  $\Omega_{l_t, d_s}$ :** In order to generate  $\Omega_{l_t, d_s}$ , the only necessary resource is the monolingual text of PBC in the target language (source domain). For embedding creation, we use fasttext (Bojanowski et al., 2017) which leverages subword information within the skip-gram architecture.

**Target embedding  $\Omega_{l_t, d_t}$ :** For generation of  $\Omega_{l_t, d_t}$ , we require a monolingual text collection in the target domain to train the embedding space. An alternative is to use pretrained embeddings in the domain of interest (e.g. Twitter or News). In particular, in our experiments, we use skip-gram embeddings, pretrained on Wikipedia for French, Macedonian, Spanish, Czech, and German (Conneau et al., 2017), as well as German, Italian, French, and Spanish pretrained embeddings on Twitter, provided by (Deriu et al., 2017; Cieliebak et al., 2017).

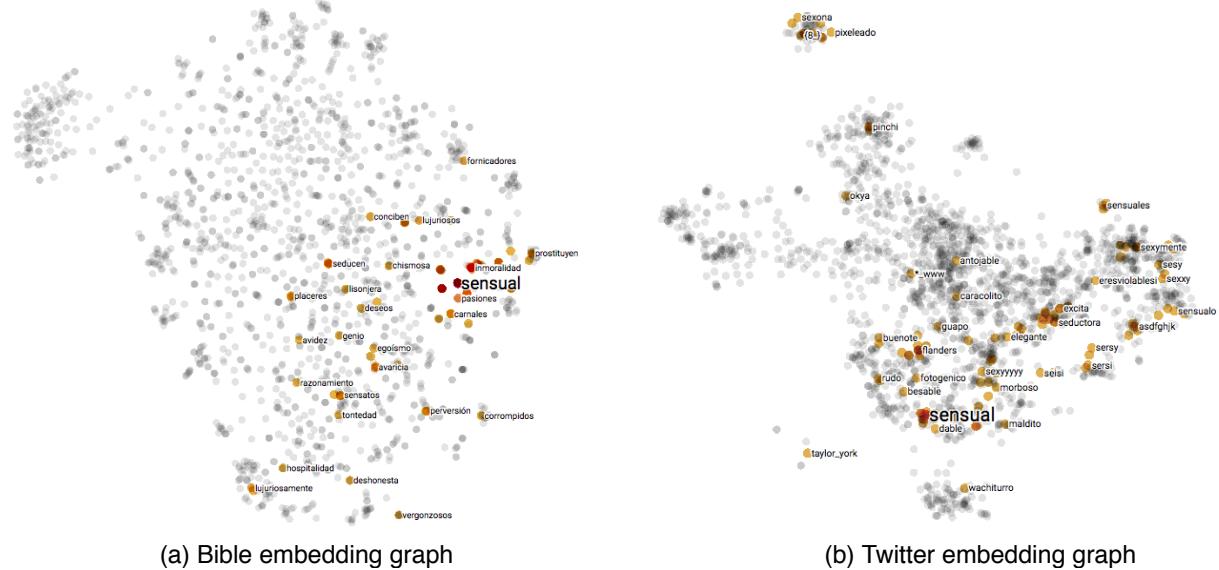


Figure 2: Neighbors of the word ‘sensual’ in Spanish, in the bible embedding graph (a) and the twitter embedding graph (b). Our unsupervised drift weighting method found this word in Spanish to be the most changing word from bible context to the twitter context. Looking more closely at the neighbors, the word sensual in the biblical context has been associated with negative sentiment of sins. However, in the twitter domain, it has a positive sentiment. This example shows how our unsupervised method can improve the quality of sentiment lexica.

### 2.3 Word sentiment classification for the evaluation of sentiment lexica

In order to evaluate *UniSent*, we use it as seed lexicon for word sentiment classification on top of embedding features. Different methods can be used to predict the sentiment of words in the target domain using embedding spaces. These include supervised methods e.g., UltraDense (Rothe et al., 2016), linear classifiers and regressors (e.g. SVM, SVR, and logistic regression) or semi-supervised methods, e.g. SentProp (Hamilton et al., 2016). In this paper, we use logistic regression for the classification model. Using the language model based embedding space as the representation has two major benefits: First, the semantic continuity of the embedding space allows for propagation of sentiment labels to a larger vocabulary. This enables the annotation of further word pairs  $(w_{l_t, d_t}, \hat{y})$ , where  $w_{l_t, d_t} \in V_{l_t, d_t}$  is a vocabulary of arbitrary size in the target domain. Secondly, using embeddings trained in a specific domain result in the implicit incorporation of semantic structures specific to the target domain (which are reflected in the embedding space), i.e. an implicit domain adaptation.

**UniSent versus confidence weighted UniSent:** For the word sentiment classification we train a logistic regression classifier with the annotated

pairs of  $(w_{l_t, d_t}, y)$  represented in the target embedding space. In order to incorporate the drift weights, we treat the weights coming from *DomDrift* as sample weights in the logistic regression model, i.e.  $(w_{l_t, d_t}, y, s_w)$ ’s are the training instances to the logistic regression classifier, where  $s_w = \frac{1}{\lambda_w}$ , is the seed weight calculated as the inverse of *DomDrift* score. Our evaluation in section §3.4 shows that this (simple) method is very effective and creates accurate resources. In the hyper-parameter tuning for logistic regression we also fine tune the exponent of this weight.

## 3 Experiments and Evaluation

### 3.1 Experimental Setup

**Select Gold Standard Data** As gold standard sentiment lexica for the evaluation of *UniSent*, we select manually created lexica in Czech (Veselovská and Bojar, 2013), German (Waltinger, 2010), French (Abdaoui et al., 2014), Macedonian (Jovanoski et al., 2016), and Spanish (Perez-Rosas et al., 2012). These lexica contain general domain words (as opposed to Twitter or Bible). As gold standard for Twitter we use the emoticon dataset in (Wiebe et al., 2005; Hogenboom et al., 2013) and perform emoticon sentiment prediction for different languages.

## 3.2 Train-test split

In order to evaluate the UniSent, here we create train-test split for training and testing the seeds created in the projection step (see Section § 2.2). We first split *UniSent* and our gold standard lexica as illustrated in Figure 3. In order to design a fair evaluation, we form our training and test sets as follows:

**(i) UniSent-Train-Lexicon:** For the evaluation of the UniSent, we use words in *UniSent* as sentiment seeds for training in the target domain; for this purpose, we use words  $w \in A \cup C$  (Figure 3).

**(ii) Manual-Train-Lexicon:** In order to obtain an upper bound for the UniSent performance, we compare the use of UniSent-Train-Lexicon against the use of words in the gold standard lexicon as sentiment seeds for the training in the target domain. For this purpose, we use words  $w \in B \cup C$  (Figure 3).

**(iii) Test-Lexicon:** we randomly exclude a set of words in the  $\{\text{Manual-Train-Lexicon} \cup B\}$ . In the selection of the sampling size, we make sure that *UniSent-Train-Lexicon* and *Manual-Train-Lexicon* would contain approximately the same number of words (Figure 3).

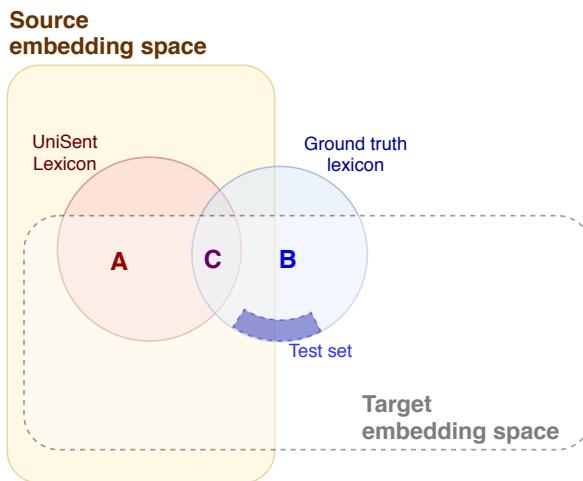


Figure 3: Data split used in the experimental setup of UniSent evaluation: Set (C) is the intersection of the target embedding space words (Wikipedia or Twitter) and the UniSent lexicon as well as the manually created lexicon. Set (A) is the intersection of the target embedding space words and the UniSent lexicon, excluding set (C). Set (B) is the intersection of the target embedding space words and the manually created lexicon, excluding set (C).

## 3.3 Evaluation

As discussed in §2.2, we use the (manually created) English sentiment lexicon (WKWSCI) in (Khoo and Johnkhan, 2018) as a resource to be projected to over 1000+ languages. We project positive and negative sentiments to create positive and negative sentiment lexica for each language.

Our evaluation of this work is two-fold. On the one hand, we evaluate the overall quality of *UniSent* by comparing it against our manually created gold standard datasets in the Wikipedia domain. Second, we investigate the influence of *DomDrift* w.r.t. the adaptation of UniSent for Wikipedia and Twitter domains.

**(i) Evaluation of UniSent vs. manually created lexica:** We compare the application of UniSent for the word sentiment classification task against the manually created lexica in the following cases: (i) choice of the most frequent sentiment, (ii) use of manually created lexicon as sentiment seeds. We use the train and test seed lexica as discussed in 3.2 for training and testing of the logistic regression on top of target embedding features. We also include the sentiment classification results using the confidence weighted version of *UniSent*, where the shift between the vocabularies of the Bible and Wikipedia are calculated by *DomDrift* and are used as sample weights in the logistic regression model.

**(ii) Comparison of UniSent vs. confidence weighted UniSent in the Twitter domain for emoticon prediction:** To show that our adaptation method also works well on domains like Twitter, we propose a second evaluation in which we use *UniSent* together with *DomDrift* to predict the sentiment of emoticons in Twitter. Since emoticons are almost language independent, we could use the same resource for the evaluation of German, Italian, French, and Spanish, where their monolingual pretrained embeddings are available for these languages (Deriu et al., 2017; Cieliebak et al., 2017). In the adaptation step, we compute the shift between the vocabularies of Bible and Twitter. We use the *UniSent* seeds for training a logistic regression model on Twitter embedding and evaluate the classifier for the Emoticon sentiment prediction. We perform this evaluation for German, Italian, French, and Spanish, where Twitter pretrained embedding is available.

Table 1: We evaluate *UniSent* against the gold standard datasets in Czech, German, French, Macedonian, and Spanish. The two last columns report the accuracy and macro-F1 (averaged F1 over positive and negative classes) of *Unisent* before and after the application of the drift weighting step. The two first columns report the performance of the baseline and manually created lexicon. Note that the baseline is constantly considering the majority label.

Language	Freq. sentiment baseline	Manual Lexicon		UniSent Lexicon		Confidence Weighted UniSent Lexicon	
		target-language-specific	acc	macro-F1	projection	acc	macro-F1
French	0.62	0.84	0.83	0.73	0.72	0.74	0.74
Macedonian	0.70	0.86	0.84	0.80	0.77	0.81	0.78
Spanish	0.64	0.82	0.80	0.78	0.76	0.80	0.77
Czech	0.62	0.87	0.87	0.82	0.81	0.79	0.78
German	0.52	0.87	0.87	0.82	0.81	0.81	0.80

Table 2: We evaluate *UniSent* using twitter emoticon dataset. We use monolingual Twitter embeddings in German, Italian, French, and Spanish. The two last columns report the accuracy and macro-F1 (averaged F1 over positive and negative classes) of *Unisent* before and after the application of the drift weighting step.

Language	Freq. sentiment baseline	UniSent Lexicon		Confidence Weighted UniSent Lexicon	
		acc	macro-F1	projection	acc
French	0.62	0.73	0.73	0.75	0.74
Spanish	0.62	0.73	0.73	0.76	0.76
German	0.62	0.80	0.79	0.80	0.79
Italian	0.62	0.76	0.76	0.75	0.75

### 3.4 Results

The evaluation results are reported in Tables 1 and 2. Table 1 compares *UniSent* and its confidence weighted version to the manually created lexica in Czech, German, French, Macedonian, and Spanish as well as a naive baseline of choosing the most frequent sentiment. For all evaluated languages, accuracy as well as macro-F1 are close to 0.8, showing that *UniSent* is a high-quality resource performing close enough to manually created seeds and clearly better than the most frequent sentiment baseline. Since in this evaluation the presented languages did not have any further advantage than having a manually created lexicon for evaluation, we can assume that *UniSent* would work within the same range of accuracy for any of the low-resource languages as long as a monolingual embedding (which is also cheap to obtain for low-resource languages) can be available for the target domain. Our drift weighting method brings gains in several languages: French, Macedonian, and Spanish.

In Table 2 we compare the quality of *UniSent* in prediction of the gold standard emoticon sentiments in the Twitter domain. The results show that (i) *UniSent* clearly outperforms the baseline of the

most frequent sentiment label and (ii) our domain adaptation technique brings small improvements for French and Spanish.

In order to illustrate the function of *DomDrift* we visualized the embedding space of biblical domain and twitter domain for the word achieving the highest drift score in Spanish, i.e., word *sensual* (Figure 2). The neighborhood of this word in both domains is shown in the figure. In Biblical texts, this word has the connotation of *sin* which has a negative polarity. But in the Twitter domain, the same word is associated with *sexy*, which has a positive polarity. This example shows that for certain pairs of domains and languages use of *DomDrift* weights in the use of sentiment lexicon can improve the performance of sentiment analysis.

### 4 Discussion and Conclusion

In this work, we introduced *UniSent* universal sentiment lexica for 1000+ languages, which is to the best of our knowledge, the largest sentiment resource to date in terms of the number of covered languages, including many low resource ones. Although *UniSent* is created based on a specific domain (bible), our evaluation of *UniSent* on Czech, German, French, Macedonian, and Span-

ish showed that it can achieve macro-F1 scores  $\approx 0.8$  in word sentiment classification, which is comparable to the use of manually annotated resources. Given that many of covered languages in *UniSent* are very low-resource, *UniSent* can be regarded among the only computational linguistic resources available for those low-resource languages making sentiment analysis possible in such low-resource setups. In addition, through accurate prediction of Twitter emoticon sentiment for German, Italian, French, and Spanish, we showed that *UniSent* can be even used in a very different target domain and still performs quite well in sentiment analysis.

Furthermore, we proposed *DomDrift*, a method to quantify domain drift for words in the *UniSent* given an embedding space in the target domain. *DomDrift* compares the neighborhood of the word in the embedding spaces of source and target domains. Incorporation of *DomDrift* scores in the use of *UniSent* for sentiment classification outperformed vanilla *UniSent* in French, Spanish, and Macedonian in the Wikipedia domain and French and Spanish in the twitter domain. Not further improving the results for German, Czech and Italian languages might be because of the sufficiency of the target embedding usage for domain adaptation (Jurafsky and Martin, 2014) in those. On the other hand, the fact that Spanish and French performances improved on both Wikipedia and Twitter domains when *DomDrift* is used, might show that the necessity of *DomDrift* can be related to certain property of the target language, which can be further explored as future work.

## References

- Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2014. Feel: French extended emotional lexicon. *ELRA Catalogue of Language Resources. ISLRN*, pages 041–639.
- Haithem Afli, Sorcha McGuire, and Andy Way. 2017. Sentiment translation for low resourced languages: Experiments on irish general election tweets. In *18th International Conference on Computational Linguistics and Intelligent Text Processing*.
- Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.
- Ehsaneddin Asgari and Mohammad R.K. Mofrad. 2016. Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (weld) as a quantitative measure of language distance. In *In Proceedings of the NAACL-HLT Workshop on Multilingual and Cross-lingual Methods in NLP, San Diego, CA*, pages 65–74. Association for Computational Linguistics.
- Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *EMNLP*.
- Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018a. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *ACL*.
- Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018b. Projecting embeddings for domain adaptation: Joint modeling of sentiment analysis in diverse domains. In *COLING*.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*.
- Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.
- George Casella and Roger L Berger. 2002. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA.
- Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland. Association for Computational Linguistics.
- Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A twitter corpus and benchmark resources for german sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51.
- Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *LREC*.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.
- Mohammad Darwich, Shahruhul Azman Mohd Noah, and Nazlia Omar. 2017. Minimally-supervised sentiment lexicon induction model: A case study of malay sentiment analysis. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, pages 225–237. Springer.
- Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. 2017. Leveraging large amounts of weakly supervised data for multilingual sentiment classification. In *Proceedings of the 26th international conference on world wide web*, pages 1045–1052. International World Wide Web Conferences Steering Committee.
- Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

- Guy Emerson, Liling Tan, Susanne Fertmann, Alexis Palmer, and Michaela Regneri. 2014. Seedling: Building and using a seed corpus for the human language project. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 77–85.
- William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 595. NIH Public Access.
- Viktor Hangya, Fabienne Braune, Alexander Fraser, and Hinrich Schütze. 2018. Two methods for domain adaptation of bilingual tasks: Delightfully simple and broadly applicable. In *ACL*.
- Alexander Hogenboom, Daniella Bal, Flavius Frasincar, Malissa Bal, Franciska de Jong, and Uzay Kaymak. 2013. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th annual ACM symposium on applied computing*, pages 703–710. ACM.
- Daniel J Hopkins and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.
- Dame Jovanoski, Veno Pachovski, and Preslav Nakov. 2016. On the impact of seed words on sentiment polarity lexicon induction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1557–1567.
- Dan Jurafsky and James H Martin. 2014. *Speech and language processing*, volume 3. Pearson London.
- Christopher SG Khoo and Sathik Basha Johnkhan. 2018. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.
- Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.
- Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. *Oceania*, 135(273):40.
- Thomas Mayer and Anders Johannsen. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Veronica Perez-Rosas, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. In *LREC*, volume 12, page 73.
- Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777, San Diego, California. Association for Computational Linguistics.
- Kateřina Veselovská and Ondřej Bojar. 2013. Czech sublex 1.0. *Charles University, Faculty of Mathematics and Physics, Institute of Formal ....*
- Ulli Waltinger. 2010. Germanpolarityclues: A lexical resource for german sentiment analysis. In *LREC*, pages 1638–1642.
- Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics.
- 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.
- Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. 2016. Quantifying political leaning from tweets, retweets, and tweeters. *IEEE transactions on knowledge and data engineering*, 28(8):2158–2172.