LowREm: A Repository of Word Embeddings for 87 Low-Resource Languages Enhanced with Multilingual Graph Knowledge

Anonymous ACL submission

Abstract

Contextualized embeddings based on large language models (LLMs) are available for various languages, but their coverage is often limited for lower resourced languages. Training LLMs for such languages is often difficult due to insufficient data and high computational cost. Especially for very low resource languages, static word embeddings thus still offer a viable alternative. There is, however, a notable lack of comprehensive repositories with such embeddings for diverse languages. To address this, we present **LowREm**, a centralized repository of static embeddings for 87 low-resource languages. We also propose a novel method to enhance GloVe-based embeddings by integrating multilingual graph knowledge, utilizing another source of knowledge, which is beneficial especially for low-resource languages. We demonstrate the superior performance of our enhanced embeddings as compared to contextualized embeddings extracted from XLM-R on sentiment analysis. Our code and data are publicly available under URL.

1 Introduction

Word embedding methods have revolutionized Natural Language Processing (NLP) by capturing semantic relationships between words using co-occurrence statistics in large text corpora (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017). This data-driven approach has significantly improved various NLP tasks (Lample et al., 2017; Xie et al., 2018; Almeida and Xexéo, 2019).

While contextual embeddings like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT (Radford et al., 2019) nowadays provide better performance than static embeddings in many tasks, their training is computationally expensive (Strubell et al., 2019) and ineffective for data-scarce languages due to their data hunger and the curse of multilinguality (Conneau et al., 2020). Also, static word embeddings remain crucial for tasks such as explaining word vector spaces (Vulić et al., 2020), bias detection and removal (Gonen and Goldberg, 2019; Manzini et al., 2019), and information retrieval (Yan et al., 2018). Existing resources for multingual embedding data bases (Ferreira et al., 2016; Grave et al., 2018) often suffer from limited scope and outdated data, potentially worsening their ability to capture the dynamic nature of language and adequately support low-resource languages. We want to fill this gap by providing **LowREm**, a large database of static word embeddings for 87 low-resource languages.

As for large language models (LLMs), the training of word embeddings suffers from the lack of high-quality data in low-resource languages (to a smaller degree). Including other types of data for improving word representations is thus beneficial especially for low-resource languages. Knowledge graphs provide such an alternative to textual knowledge, with rich semantic and multilingual sources of information, including synonyms, antonyms, morphological forms, definitions, etimological relations, translations, and more (Miller, 1995; Speer et al., 2017; Navigli and Ponzetto, 2012). Such structured and cross-lingual information can be used to improve the quality of classical word representations (Faruqui et al., 2014; Sakketou and Ampazis, 2020), which are only trained on co-occurence statistics.

To that end, we propose a new simple yet effective method for including graph information into word embeddings based on Mikolov et al. (2013b). We learn a projection matrix from static embeddings to a combined space, effectively overcoming the limitations of retrofitting, which only enhances a limited vocabulary.

In summary, our contributions in this work are two-fold: First, we present **LowREm**, a centralized resource of static word embeddings for low-resource languages, specifically focusing on word embeddings trained with GloVe (Penning-

ton et al., 2014). Second, we propose an effective method to improve embeddings by incorporating more knowledge in the form of multilingual knowledge graphs, which is especially important for low-resource languages, where resources are usually very scarce.

2 Related Work

We briefly describe the most prominent graph knowledge sources, word embeddings, and existing methods for improving embeddings with graphs.

Graph knowledge sources. Among most used knowledge graphs for natural language are WordNet (Miller, 1995) and BabelNet (Navigli and Ponzetto, 2012). Wordnet is a lexical database that organizes English words into sets of synonyms called synsets, providing short definitions and usage examples. BabelNet is a multilingual encyclopedic dictionary and semantic network, which integrates lexicographic and encyclopedic knowledge from WordNet, Wikipedia, etc., focused on named entities. In our work, we use ConceptNet (Speer et al., 2017), a multilingual, domain-general knowledge graph that connects words and phrases from various natural languages with labeled, weighted edges representing relationships between terms. Unlike other knowledge graphs, ConceptNet is not a monolingual collection of named entities but focuses on commonly used words and phrases across multiple languages.

Word embeddings. *Word2Vec* (Mikolov et al., 2013a) uses shallow neural networks to produce word vectors. It comes in two types: Continuous Bag of Words (CBOW) and Skip-gram. *GloVe* (Global Vectors for Word Representation) (Pennington et al., 2014) word embeddings are created by aggregating global word-word co-occurrence statistics from a corpus. The resulting vectors capture both local and global semantic relationships, providing robust word representations that outperform many alternatives in various NLP tasks. *FastText* (Bojanowski et al., 2017) extends Word2Vec by representing words as bags of character n-grams, capturing subword information and handling out-of-vocabulary words more effectively. *Numberbatch*, part of the ConceptNet project (Speer et al., 2017), is a set of word embeddings that integrates knowledge from ConceptNet with distributional semantics from GloVe and Word2Vec. Numberbatch uses a retrofitting approach (Faruqui et al., 2014) to enhance embeddings with structured semantic knowledge. Retrofitting often results in a limited vocabulary for underrepresented languages (Speer and Lowry-Duda, 2017).

Improving Embeddings with Knowledge Graphs. There are various methods to improve word embeddings by incorporating external knowledge graphs or semantic networks (Dieudonat et al., 2020). *Retrofitting* (Faruqui et al., 2014) is a post-processing technique that adjusts pre-trained word embeddings using information from knowledge graphs or semantic lexicons. The key idea is to infer new vectors that are close to their original embeddings while also being close to their neighbors in the graph or lexicon. *Expanded retrofitting* (Speer et al., 2017), used for ConceptNet Numberbatch, optimizes over a larger vocabulary including terms from the knowledge graph not present in the original embeddings, but it still does not retrofit all the words in the original embedding space. Other existing methods that integrate contextualized embeddings with knowledge graph embeddings often use attention mechanisms, as demonstrated by works such as Peters et al. (2019) and Zhang et al.(2019). These methods specifically enhance BERT embeddings by incorporating external knowledge bases.

3 Method

We propose a method for merging GloVe embeddings with graph-based embeddings derived from ConceptNet knowledge, while preserving the vocabulary size of GloVe, following two steps: First, we use Singular Value Decomposition (SVD) on concatenated word embeddings from GloVe and PPMI-based graph embeddings to generate a shared embedding space. We do so for the part of the vocabulary that is shared between GloVe and the knowledge graph. Second, we learn a linear transformation from GloVe into this joined space to obtain embeddings for all words in the original GloVe vocabulary.

3.1 GloVe Embeddings

We trained GloVe embeddings using the original C code[1]. The model was trained by stochastically sampling nonzero elements from the co-occurrence matrix X, over 100 iterations, to pro-

[1] All embeddings are open-sourced on HuggingFace.

duce 300-dimensional vectors. We used a context window of ten words to the left and ten words to the right. Words with fewer than 5 co-occurrences were excluded for languages with over 1 million tokens in the training data, and the threshold was set to 2 for languages with smaller datasets. We used data from CC100[2] (Wenzek et al., 2020; Conneau et al., 2020) for training the static word embeddings. We set $x_{max} = 100$, $\alpha = \frac{3}{4}$, and use AdaGrad optimization (Duchi et al., 2011) with an initial learning rate of 0.05.

3.2 Graph Embeddings

To build ConceptNet-based word embeddings, we follow the method used for constructing ConceptNet Numberbatch embeddings (Speer et al., 2017). We represent the ConceptNet graph as a sparse, symmetric term-term matrix, where each cell is the sum of the occurences of all edges connecting the two terms. Unlike the original method, we do not discard terms connected to fewer than three edges, as we deal with low-resource langauges.

We calculate embeddings from this matrix by applying pointwise mutual information (PMI) with context distributional smoothing set to 0.75, clipping negative values to yield positive PMI (PPMI), which follows practical recommendations by (Levy et al., 2015). We then reduce the dimensionality to 300 dimensions using truncated SVD and combine terms and contexts symmetrically to form a single matrix of word embeddings, called ConceptNet-PPMI. This matrix captures the overall graph structure of ConceptNet.

We computed ConceptNet-PPMI embeddings for the entire ConceptNet data, covering 304 languages, which we call *PPMI (All)*. Additionally, we constructed separate graph embedding spaces, *PPMI (Single)*, for each specific language, using only the portion of ConceptNet data for that language. This approach was adopted because the initial co-occurence matrices for individual languages are less sparse while still being multilingual in nature.

3.3 Singular Value Decomposition (SVD)

We first concatenate GloVe and ConceptNet-PPMI vectors for all words that are in the shared vocabulary, resulting in 600 dimensional vectors[3]. Afterwards, we reduce the dimensionality and remove some of the variance coming from redundant features. The matrix M representing merged GloVe and ConceptNet-PPMI can be approximated with a truncated SVD:

$$M \approx U \Sigma V^T$$

where Σ is truncated to a $k' \times k'$ diagonal matrix of the k' largest singular values, and U and V are correspondingly truncated to have only these k' columns. U can then be used as a matrix mapping the original vocabulary to a smaller set of features[4].

3.4 Linear Transformation

To obtain embeddings for the entire vocabulary from the original GloVe embedding space (i.e. not only the common words), we find a linear projection matrix between the spaces and project the GloVe embeddings onto the merged embedding space, similar to Mikolov et al. (2013c), using a gradient descent optimization on a linear regression model.

4 Experiments

In this section, we describe the chosen languages, tasks for measuring the effectiveness of the proposed method, and conducted experiments.

4.1 Languages

We train GloVe word embeddings for 87 languages from CC100 (Wenzek et al., 2020) categorized as low-resource from class 3 to 0 based on the classification by Joshi et al. (2020). Additionally, we generated graph embeddings for 72 languages present in both CC100 and ConceptNet. We applied our merging method to enhance the quality of the original embeddings for these languages. Details on the languages are specified in Table 2 of the Appendix.

4.2 Embedding Evaluation

For evaluating the embeddings, we perform an extrinsic evaluation using a downstream NLP task - sentiment analysis (SA). Obtaining intrinsic evaluation datasets for most underresourced languages is challenging. We use a self-gathered collection of datasets for diverse languages. Details on data sources and distribution for SA datasets are available in Table 3 of the Appendix. For the imbalanced datasets (Swahili, Nepali, Uyghur, Latvian,

[2] https://huggingface.co/datasets/cc100

[3] ConceptNet-PPMI embeddings were normalized to be in the range of the Glove embeddings

[4] We dismiss the weighting of U by the singular values from Σ, which was noted to work better for semantic tasks (Levy et al., 2015)

Slovak, Slovenian, Uzbek, Bulgarian, Yoruba, Bengali, Hebrew, Telugu), we used random undersampling to balance the data distribution.

4.3 Experimental Setup

For evaluation, we use a Support Vector Machine (SVM, Boser et al. (1992)) to predict sentiment. Our experiments involved training the SVM classifier using GloVe embeddings, GloVe+PPMI (Single) and GloVe+PPMI (All) to represent sentences while fitting the data to the model. Sentence embeddings were obtained by summing up embeddings for individual words extracted from the corresponding dictionaries, and the SVM was trained on top of these representations. We fixed the regularization parameter C to 1.0 and used the RBF kernel to reduce the influence of the hyperparameters on the resulting scores. We use macro F1 score to measure performance. As baseline, we use XLM-R (Conneau et al., 2020) for obtaining sentence embeddings by summing up the last hidden states of the model and consequently train an SVM classifier on these representations.

Lang	XLM-R	G	G+P (Single)	G+P (All)
am	0.616	**0.881**	0.86	0.88
su	0.674	0.798	**0.822**	0.812
sw	0.473	0.68	0.701	**0.714**
si	0.631	0.848	0.85	**0.857**
ka	0.495	0.861	**0.87**	0.861
ne	0.542	0.643	0.674	**0.688**
ug	0.386	0.746	**0.811**	**0.811**
yo	0.52	0.721	0.709	**0.738**
ur	0.526	0.676	**0.746**	0.745
mk	0.351	**0.716**	0.711	0.7
mr	0.809	0.903	**0.905**	0.902
bn	0.551	0.875	**0.881**	0.878
te	0.603	0.806	0.808	**0.817**
uz	0.64	**0.808**	0.806	0.806
az	0.6	0.744	**0.746**	0.745
bg	0.568	0.786	0.801	**0.805**
sl	0.582	0.749	0.779	**0.788**
lv	0.606	0.783	**0.787**	**0.787**
sk	0.657	0.756	**0.806**	0.805
ro	0.622	0.805	**0.85**	0.847
he	0.672	0.788	**0.824**	0.822
cy	0.588	0.77	0.789	**0.801**
da	0.77	0.863	**0.908**	0.903

Table 1: Macro Average F1 Scores for SA per language, for XLM-R, GloVe (G), GloVe + PPMI (G+P), Single and All. Maximum per row in **bold**.

4.4 Results

We evaluate the performance of the proposed GloVe+PPMI embeddings on sentiment analysis (SA) tasks for 23 low-resource languages. Table 1 presents macro average F1 scores for SA. The GloVe+PPMI (Single) and GloVe+PPMI (All) embeddings consistently outperform the original GloVe embeddings across most languages. We observed fine improvements even for the languages with a small number of common vocabulary between GloVe and PPMI, such as Uygur, Sundanese and others (more details on vocabulary overlap in Section C of the Appendix).

Our method especially also outperforms XLM-R-base embeddings, highlighting the potential of static embeddings enhanced with multilingual graph knowledge in low-resource settings.

Overall, the results indicate that integrating graph-based knowledge into GloVe embeddings through PPMI largely improves the performance of the embeddings on SA tasks. The consistent improvement in F1 scores across various languages suggests that the additional semantic and multilingual relationships captured by the graph-based approach provide valuable context that the original GloVe embeddings lack. This is particularly beneficial for low-resource languages where the amount of available training data is limited.

5 Conclusion

In this work, we addressed a need for baseline word embeddings in low-resource languages by creating a centralized resource of pre-trained static embeddings for 87 diverse languages. Our novel method integrates GloVe embeddings with graph-based knowledge from ConceptNet using Singular Value Decomposition (SVD) and a linear transformation to merge the embedding spaces. This approach enhances the original embeddings, as demonstrated by superior performance on SA tasks across various languages compared to both GloVe and contextualized embeddings extracted from XLM-R.

Our contributions include not only the proposed method but also the provision of **LowREm**, an extensive repository of GloVe word embeddings, accessible for a wide range of low-resource languages. This resource is aimed to support and advance NLP applications and research in underrepresented languages, ensuring that the benefits of modern NLP techniques extend to all linguistic communities.

Limitations

While our contribution lies in providing baseline models across a wide range of languages, there are

several limitations to consider. First, we acknowledge that our evaluation was focused on extrinsic task of sentiment analysis, and we did not extensively evaluate both GloVe and enhanced GloVe embeddings on intrinsic tasks due to a lack of corresponding datasets. Future work could involve the evaluation of our method on existing intrinsic evaluation datasets and the creation of such resources for low-resource languages, allowing for a more comprehensive assessment of the quality and performance of the embeddings. Secondly, the quantity and quality of training data remain important factors influencing the effectiveness of word embeddings. Despite efforts to leverage large-scale corpora such as CC100 and ConceptNet, there are still limitations in the availability and diversity of training data, particularly for low-resource languages. Furthermore, while our method of merging GloVe embeddings with graph-based embeddings has shown promising results, there is potential for further refinement and exploration of alternative merging and projection techniques. Future research could investigate advanced fusion and projection methods, potentially leading to more enhanced representations for low-resource languages.

Acknowledgments

References

Felipe Almeida and Geraldo Xexéo. 2019. Word embeddings: A survey. *ArXiv*, abs/1901.09069.

Adam Amram, Anat Ben David, and Reut Tsarfaty. 2018. Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from Modern Hebrew. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2242–2252, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Jože Bučar, Martin Žnidaršič, and Janez Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52(3):895–919.

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

Piyumal Demotte, Lahiru Senevirathe, Binod Karunanayake, Udyogi Munasinghe, and Surangika Ranathunga. 2020. Sentiment analysis of sinhala news comments using sentence-state lstm networks. pages 283–288.

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

Lea Dieudonat, Kelvin Han, Phyllicia Leavitt, and Esteban Marquer. 2020. Exploring the combination of contextual word embeddings and knowledge graph embeddings. *ArXiv*, abs/2004.08371.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Luis Espinosa-Anke, Geraint Palmer, Padraig Corcoran, Maxim Filimonov, Irena Spasić, and Dawn Knight. 2021. English–welsh cross-lingual embeddings. *Applied Sciences*, 11(14):6541.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

Daniel C. Ferreira, André F. T. Martins, and Mariana S. C. Almeida. 2016. Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2019–2028, Berlin, Germany. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning

word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tim Isbister, Fredrik Carlsson, and Magnus Sahlgren. 2021. Should we stop training more monolingual models, and simply use machine translation instead? In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 385–390, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Dame Jovanoski, Veno Pachovski, and Preslav Nakov. 2015. Sentiment analysis in Twitter for Macedonian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 249–257, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Muhammad Yaseen Khan, Shah Muhammad Emaduddin, and Khurum Nazir Junejo. 2017. Harnessing english sentiment lexicons for polarity detection in urdu tweets: A baseline approach. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 242–249. IEEE.

Muhammad Yaseen Khan and Muhammad Suffian Nizami. 2020. Urdu sentiment corpus (v1.0): Linguistic exploration and visualization of labeled datasetfor urdu sentiment analysis. In *2020 IEEE 2nd International Conference On Information Science Communication Technology (ICISCT)*. IEEE.

Elmurod Kuriyozov, Sanatbek Matlatipov, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2019. Construction and evaluation of sentiment datasets for low-resource languages: The case of uzbek. In *Human Language Technology. Challenges for Computer Science and Linguistics - 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17-19, 2019, Revised Selected Papers*, volume 13212 of *Lecture Notes in Computer Science*, pages 232–243. Springer.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Siyu Li, Kui Zhao, Jin Yang, Xinyun Jiang, Zhengji Li, and Zicheng Ma. 2022. Senti-exlm: Uyghur enhanced sentiment analysis model based on xlm. *Electronics Letters*, 58(13):517–519.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

LocalDoc. 2024. Sentiment analysis dataset for Azerbaijani.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and R. Mamidi. 2022a. Multi-task text classification using graph convolutional networks for large-scale low resource language. *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022b. Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–34.

Antonio Martínez-García, Toni Badia, and Jeremy Barnes. 2021. Evaluating morphological typology in zero-shot cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3136–3153, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *ArXiv*, abs/1309.4168.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013c. Exploiting similarities among languages for machine translation. *ArXiv*, abs/1309.4168.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Shamsuddeen Muhammad, Idris Abdulmumin, Abinew Ayele, Nedjma Ousidhoum, David Adelani, Seid Yimam, Ibrahim Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Felermino Ali, Davis David, Salomey Osei, Bello Shehu-Bello, Falalu Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Messelle, Hailu Balcha, Sisay Chala, Hagos Gebremichael, Bernard Opoku, and Stephen Arthur. 2023a. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 task 12: Sentiment analysis for African languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2319–2337, Toronto, Canada. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Samuel Pecar, Marian Simko, and Maria Bielikova. 2019. Improving sentiment classification in Slovak language. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 114–119, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. L3cube-mahasent-md: A multi-domain marathi sentiment analysis dataset and transformer models. In *Pacific Asia Conference on Language, Information and Computation*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Flora Sakketou and Nicholas Ampazis. 2020. A constrained optimization algorithm for learning glove embeddings with semantic lexicons. *Knowledge-Based Systems*, 195:105628.

Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource Bengali language. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 50–60, Online. Association for Computational Linguistics.

Oyesh Mann Singh, Sandesh Timilsina, Bal Krishna Bal, and Anupam Joshi. 2020. Aspect based abusive sentiment detection in nepali social media texts. *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 301–308.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Robyn Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Vancouver, Canada. Association for Computational Linguistics.

Uga Sproģis and Matīss Rikters. 2020. What Can We Learn From Almost a Decade of Food Tweets. In *In Proceedings of the 9th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2020)*, Kaunas, Lithuania.

Nicolas Stefanovitch, Jakub Piskorski, and Sopho Kharazi. 2022. Resources and experiments on sentiment classification for Georgian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1613–1621, Marseille, France. European Language Resources Association.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Anca Tache, Gaman Mihaela, and Radu Tudor Ionescu. 2021. Clustering word embeddings with self-organizing maps. application on LaRoSeDa - a large Romanian sentiment data set. In *Proceedings of the*

16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 949–956, Online. Association for Computational Linguistics.

Tarikwa Tesfa, Befikadu Belete, Samuel Abera, Sudhir Kumar Mohapatra, and Tapan Kumar Das. 2024. Aspect-based sentiment analysis on amharic text for evaluating ethio-telecom services. In *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, pages 1–6.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Fengqi Yan, Qiaoqing Fan, and Mingming Lu. 2018. Improving semantic similarity retrieval with word embeddings. *Concurrency and Computation: Practice and Experience*, 30(23):e4489.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Appendix

A Language Details

ISO code	Language	Size	Class	ConceptNet	ISO code	Language	Size	Class	ConceptNet
ss	Swati	86K	1	✘	**sc**	Sardinian	143K	1	✓
yo	Yoruba	1.1M	2	✓	**gn**	Guarani	1.5M	1	✓
qu	Quechua	1.5M	1	✓	**ns**	Northern Sotho	1.8M	1	✘
li	Limburgish	2.2M	1	✓	**ln**	Lingala	2.3M	1	✓
wo	Wolof	3.6M	2	✓	**zu**	Zulu	4.3M	2	✓
rm	Romansh	4.8M	1	✓	**ig**	Igbo	6.6M	1	✘
lg	Ganda	7.3M	1	✘	**as**	Assamese	7.6M	1	✘
tn	Tswana	8.0M	2	✘	**ht**	Haitian	9.1M	2	✓
om	Oromo	11M	1	✘	**su**	Sundanese	15M	1	✓
bs	Bosnian	18M	3	✘	**br**	Breton	21M	1	✓
gd	Scottish Gaelic	22M	1	✓	**xh**	Xhosa	25M	2	✓
mg	Malagasy	29M	1	✓	**jv**	Javanese	37M	1	✓
fy	Frisian	38M	0	✓	**sa**	Sanskrit	44M	2	✓
my	Burmese	46M	1	✓	**ug**	Uyghur	46M	1	✓
yi	Yiddish	51M	1	✓	**or**	Oriya	56M	1	✓
ha	Hausa	61M	2	✓	**la**	Lao	63M	2	✓
sd	Sindhi	67M	1	✓	**ta_rom**	Tamil Romanized	68M	3	✘
so	Somali	78M	1	✓	**te_rom**	Telugu Romanized	79M	1	✘
ku	Kurdish	90M	0	✓	**pu/pa**	Punjabi	90M	2	✓
ps	Pashto	107M	1	✓	**ga**	Irish	108M	2	✓
am	Amharic	133M	2	✓	**ur_rom**	Urdu Romanized	141M	3	✘
km	Khmer	153M	1	✓	**uz**	Uzbek	155M	3	✓
bn_rom	Bengali Romanized	164M	3	✘	**ky**	Kyrgyz	173M	3	✓
my_zaw	Burmese (Zawgyi)	178M	1	✘	**cy**	Welsh	179M	1	✓
gu	Gujarati	242M	1	✓	**eo**	Esperanto	250M	1	✓
af	Afrikaans	305M	3	✓	**sw**	Swahili	332M	2	✓
mr	Marathi	334M	2	✓	**kn**	Kannada	360M	1	✓
ne	Nepali	393M	1	✓	**mn**	Mongolian	397M	1	✓
si	Sinhala	452M	0	✓	**te**	Telugu	536M	1	✓
la	Latin	609M	3	✓	**be**	Belarussian	692M	3	✓
tl	Tagalog	701M	3	✘	**mk**	Macedonian	706M	1	✓
gl	Galician	708M	3	✓	**hy**	Armenian	776M	1	✓
is	Icelandic	779M	2	✓	**ml**	Malayalam	831M	1	✓
bn	Bengali	860M	3	✓	**ur**	Urdu	884M	3	✓
kk	Kazakh	889M	3	✓	**ka**	Georgian	1.1G	3	✓
az	Azerbaijani	1.3G	1	✓	**sq**	Albanian	1.3G	1	✓
ta	Tamil	1.3G	3	✓	**et**	Estonian	1.7G	3	✓
lv	Latvian	2.1G	3	✓	**ms**	Malay	2.1G	3	✓
sl	Slovenian	2.8G	3	✓	**lt**	Lithuanian	3.4G	3	✓
he	Hebrew	6.1G	3	✓	**sk**	Slovak	6.1G	3	✓
el	Greek	7.4G	3	✓	**th**	Thai	8.7G	3	✓
bg	Bulgarian	9.3G	3	✓	**da**	Danish	12G	3	✓
uk	Ukrainian	14G	3	✓	**ro**	Romanian	16G	3	✓
id	Indonesian	36G	3	✘					

Table 2: Details of the reproduced CC-100 corpus available on HuggingFace, including languages with their ISO codes, data set sizes, low-resource classifications, and language availability in the ConceptNet knowledge graph.

B SA Data Details

Language	ISO code	Source	#pos	#neg	#train	#val	#test
Sundanese	su	Winata et al., 2023	378	383	381	76	304
Amharic	am	Tesfa et al., 2024	487	526	709	152	152
Swahili	sw	Muhammad et al., 2023a; Muhammad et al., 2023b	908	319	738	185	304
Georgian	ka	Stefanovitch et al., 2022	765	765	1080	120	330
Nepali	ne	Singh et al., 2020	680	1019	1189	255	255
Uyghur	ug	Li et al., 2022	2450	353	1962	311	530
Latvian	lv	Sproģis and Rikters, 2020	1796	1380	2408	268	500
Slovak	sk	Pecar et al., 2019	4393	731	3560	522	1042
Sinhala	si	Demotte et al., 2020	2487	2516	3502	750	751
Slovenian	sl	Bučar et al., 2018	1665	3337	3501	750	751
Uzbek	uz	Kuriyozov et al., 2019	3042	1634	3273	701	702
Bulgarian	bg	Martínez-García et al., 2021	6652	1271	5412	838	1673
Yoruba	yo	Muhammad et al., 2023a; Muhammad et al., 2023b	6344	3296	5414	1327	2899
Urdu	ur	Maas et al., 2011; Khan et al., 2017; Khan and Nizami, 2020	5562	5417	7356	1812	1812
Macedonian	mk	Jovanoski et al., 2015	3041	5184	6557	729	939
Danish	da	Isbister et al., 2021	5000	5000	7000	1500	1500
Marathi	mr	Pingle et al., 2023	5000	5000	8000	1000	1000
Bengali	bn	Sazzed, 2020	8500	3307	8264	1771	1772
Hebrew	he	Amram et al., 2018	8497	3911	8932	993	2483
Romanian	ro	Tache et al., 2021	7500	7500	10800	1200	3000
Telugu	te	Marreddy et al., 2022b; Marreddy et al., 2022a	9488	6746	11386	1634	3214
Welsh	cy	Espinosa-Anke et al., 2021	12500	12500	17500	3750	3750
Azerbaijani	az	LocalDoc, 2024	14000	14000	19600	4200	4200

Table 3: Sentiment Analysis Data Details

C Vocabulary Details

Language	SA Vocabulary Coverage (%)	Common Vocabulary
am	78	1,105
su	78	1,236
sw	88	6,425
si	89	943
ka	97	17,869
ne	78	2,650
ug	88	764
yo	22	558
ur	63	4,662
mk	83	21,692
mr	84	3,211
bn	67	3,962
te	86	12,563
uz	71	3,229
az	60	7,215
bg	84	92,436
sl	92	45,153
lv	87	17,450
sk	85	14,694
ro	90	25,704
he	90	16,032
cy	52	7,774
da	75	38,095

Table 4: Percentage of Glove and Glove+PPMI Vocabulary Coverage of SA data, Common Vocabulary Between GloVe and PPMI Embedding Spaces.