Investigating an approach for low resource language dataset creation, curation and classification: Setswana and Sepedi

Vukosi Marivate[1,2], Tshephisho Sefara[2], Vongani Chabalala[3], Keamogetswe Makhaya[4], Tumisho Mokgonyane[5], Rethabile Mokoena[6], Abiodun Modupe[7,1]
University of Pretoria[1], CSIR[2], University of Zululand[3], University of Cape Town[4],
University of Limpopo[5], North-West University[6], University of the Witwatersrand[7]
vukosi.marivate@cs.up.ac.za, tsefara@csir.co.za

Abstract

The recent advances in Natural Language Processing have only been a boon for well represented languages, negating research in lesser known global languages. This is in part due to the availability of curated data and research resources. One of the current challenges concerning low-resourced languages are clear guidelines on the collection, curation and preparation of datasets for different use-cases. In this work, we take on the task of creating two datasets that are focused on news headlines (i.e short text) for Setswana and Sepedi and the creation of a news topic classification task from these datasets. In this study, we document our work, propose baselines for classification, and investigate an approach on data augmentation better suited to low-resourced languages in order to improve the performance of the classifiers.

1. Introduction

The most pressing issues with regard to low-resource languages are the lack of sufficient language resources, like features related to automation. In this study, we introduce an investigation of a low-resource language that provides automatic formulation and customisation of new capabilities from existing ones. While there are more than six thousand languages spoken globally, the availability of resources among each of those are extraordinarily unbalanced (Nettle, 1998). For example, if we focus on language resources annotated on the public domain, as of November 2019, AG corpus released about $496,835$ news articles related to the English language from more than 200 sources[1]. Additionally, the Reuters News Dataset (Lewis, 1997) comprise roughly $10,788$ annotated texts from the Reuters financial newswire. Moreover, the New York Times Annotated Corpusholds over 1.8 million articles (Sandhaus, 2008). Lastly, Google Translate only supports around 100 languages (Johnson et al., 2017). significant amount of knowledge exists for only a small number of languages, neglecting 17% out of the world's language categories labelled as low-resource, and there are currently no standard annotated tokens in low-resource languages (Strassel and Tracey, 2016). This in turn, makes it challenging to develop various mechanisms and tools used for Natural Language Processing (NLP).

In South Africa, most of the news websites (private and public) are published in English, despite there being 11 official languages (including English). In this paper, we list the premium newspapers by circulation as per the first Quarter of 2019 (Bureau of Circulations, 2019) (Table 1). Currently, there is a lack of information surrounding 8 of the 11 official South African languages, with the exception of English, Afrikaans and isiZulu which contain most of the reported datasets. n this work, we aim to provide a general framework for two of the 11 South African languages, to create an annotated linguistic resource for Setswana and Sepedi news headlines. In this study, we applied data sources of the news headlines from the South African Broadcast Corporation (SABC) [2], their social media streams and a few acoustic news. Unfortunately, at the time of this study, we did not have any direct access to news reports, and hopefully this study can promote collaboration between the national broadcaster and NLP researchers.

Table 1: Top newspapers in South Africa with their languages

Paper	Language	Circulation
Sunday Times	English	260132
Soccer Laduma	English	252041
Daily Sun	English	141187
Rapport	Afrikaans	113636
Isolezwe	isiZulu	86342
Sowetan	English	70120
Isolezwe ngeSonto	isiZulu	65489
Isolezwe ngoMgqibelo	isiZulu	64676
Son	Afrikaans	62842

The rest of the work is organized as follows. Section 2. discusses prior work that has gone into building local corpora in South Africa and how they have been used. Section 3. presents the proposed approach to build a local news corpora and annotating the corpora with categories. From here, we focus on ways to gather data for vectorization and building word embeddings (needing an expanded corpus). We also release and make pre-trained word embeddings for 2 local languages as part of this work (Marivate and Sefara, 2020a). Section 4. investigate building classification models for the Setswana and Sepedi news and improve those classifiers using a 2 step augmentation approach inspired by work on hierarchical language models (Yu et al., 2019). Finally, Section 5. concludes and proposes a path forward for this work.

[1] http://groups.di.unipi.it/~gulli

[2] http://www.sabc.co.za/

2. Prior Work

Creating sizeable language resources for low resource languages is important in improving available data for study (Zoph et al., 2016) and cultural preservation. Focusing on the African continent, we note that there are few annotated datasets that are openly available for Natural Language Processing tasks such as classification. In South Africa, the South African Center for Digital Language Resources (SADiLaR) [3] has worked to curate datasets of local South African languages. There remain gaps such as accessing large corpora and data from sources such as broadcasters and news organizations as they have sizeable catalogs that are yet to make it into the public domain. In this work, we work to fill such a gap by collecting, annotating and training classifier models for news headlines in Setswana and Sepedi. As the data that we do find publicly is still small, we also have to deal with the challenges of Machine Learning on small data. Machine learning systems perform poorly in presence of small training sets due to overfitting. To avoid this problem, data augmentation can be used. The technique is well known in the field of image processing (Cubuk et al., 2019). Data augmentation refers to the augmentation of the training set with artificial, generated, training examples. This technique is used less frequently in NLP but a number of few studies applied data augmentation.

Silfverberg et al. (2017) use data augmentation to counteract overfitting where recurrent neural network (RNN) Encoder-Decoder is implemented specifically geared toward a low-resource setting. Authors apply data augmentation by finding words that share word stem for example **fizzle** and **fizzling** share **fizzl**. Then authors replace a stem with another string.

Zhang et al. (2015) apply data augmentation by using synonyms as substitute words for the original words. However, Kobayashi (2018) states that synonyms are very limited and the synonym-based augmentation cannot produce numerous different patterns from the original texts. Hence, Kobayashi (2018) proposes contextual data augmentation by replacing words that are predicted by a language model given the context surrounding the original words to be augmented.

As Wei and Zou (2019) states that these techniques are valid, they are not often used in practice because they have a high cost of implementation relative to performance gain. They propose an easy data augmentation as techniques for boosting performance on text classification tasks. These techniques involve synonym replacement, random insertion, random swap, and random deletion of a word. Authors observed good performance when using fraction of the dataset (%):1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, as the data size increases, the accuracy also increases for augmented and original data. Original data obtained highest accuracy of 88.3% at 100% data size while augmented data obtained accuracy of 88.6% at 50% data size.

[3] www.sadilar.org

3. Developing news classification models for Setswana and Sepedi

In this work, we investigate the development of a 2 step text augmentation method in order to be improve classification models for Setswana and Sepedi. To do this we had to first identify a suitable data source. Collect the data, and then annotate the datasets with news categories. After the data is collected and annotated, we then worked to create classification models from the data as is and then use a word embedding and document embedding augmentation approach. In this section discuss how data was collected as well as the approach we use to build classification models.

3.1. Data Collection, Cleaning and Annotation

Before we can train classification models, we first have to collect data for 2 distinct processes. First, we present our collected news dataset as well as its annotation. We then discuss how we collected larger datasets for better vectorisation.

3.1.1. News data collection and annotation

The news data we collected is from the SABC[4] Facebook pages. The SABC is the public broadcaster for South Africa. Specifically, data was collected from Motswedi ng FM (An SABC Setswana radio station) [5] and Thobela FM (An SABC Sepedi radio station) [6]. We scraped the news headlines that are published as posts on both Facebook pages. We claim no copyright for the content but used the data for research purposes. We summarize the datasets in Table 2. We visualize the token distributions in Setswana and Sepedi in Figures 1 and 2 respectively.

Table 2: News Data Sets

	Setswana	Sepedi
Corpus Size (headlines)	219	491
Number of Tokens (words)	1561	3018

Figure 1: Setswana Wordcloud

[4] http://www.sabc.co.za/

[5] https://www.facebook.com/MotswedingFM/

[6] https://www.facebook.com/thobelafmyaka/

Figure 2: Sepedi Wordcloud

As can be seen, the datasets are relatively small and as such, we have to look at other ways to build vectorizers that can better generalize as the word token diversity would be very low.

We annotated the datasets by categorizing the news headlines into: *Legal*, *General News*,*Sports*, *Other*, *Politics*, *Traffic News*, *Community Activities*, *Crime*, *Business* and *Foreign Affairs*. Annotation was done after reading the headlines and coming up with categories that fit both datasets. We show the distribution of the labels in both the Setswana and Sepedi data sets in Figures 3 and 4 respectively. For this work, we only explore single label categorization for each article. It remains future work to look at the multi-label case. As such, there might be some noise in the labels. Examples from the Sepedi annotated news corpus are shown next:

> *Tsela ya N1 ka Borwa kgauswi le Mantsole Weighbridge ka mo Limpopo ebe e tswaletswe lebakanyana ka morago ga kotsi yeo e hlagilego.*
> **Traffic**

> *Tona ya toka Michael Masutha,ore bahlankedi ba kgoro ya ditirelo tsa tshokollo ya bagolegwa bao ba tateditswego dithieeletsong tsa khomisene ya go nyakisisa mabarebare a go gogwa ga mmuso ka nko,ba swanetse go hlalosa gore ke ka lebaka la eng ba sa swanelwa go fegwa mesomong*
> **Legal**

The full dataset is made available online (Marivate and Sefara, 2020b) for further research use and improvements to the annotation[7]. As previously discussed, we used larger corpora to create language vectorizers for downstream NLP tasks. We discuss this next.

3.1.2. Vectorizers

Before we get into the annotated dataset, we needed to create pre-trained vectorizers in order to be able to build more classifiers that generalize better later on. For this reason we collected different corpora for each language in such as

[7] https://zenodo.org/record/3668495

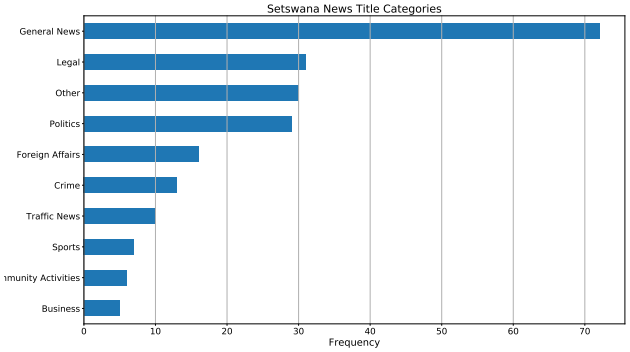

Figure 3: Setswana news title category distribution

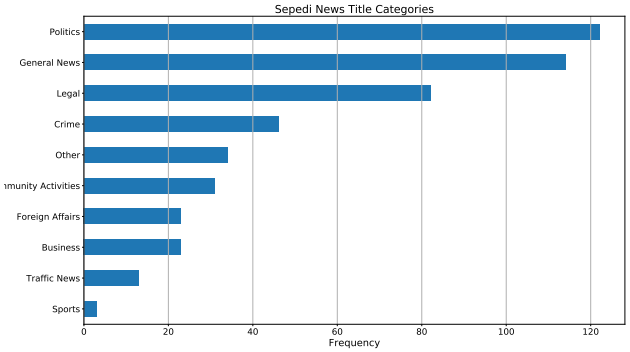

Figure 4: Sepedi news title category distribution

way that we could create Bag of Words, TFIDF, Word2vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017) vectorizers (Table 3). We also make these vectorizers available for other researchers to use.

Table 3: Vectorizer Corpora Sizes in number of lines (number of tokens)

Source	Setswana	Sepedi
Wikipedia	478(*21924*)[8]	300(*10190*)[9]
JW300[10]	874464(*70251*)	618275(*53004*)
Bible	31102(*42233*)	29723(*38709*)
Constitution[11]	7077(*3940*)	6564(*3819*)
SADILAR[12]	33144(*61766*)	67036(*87838*)
Total	**946264(*152027*)**	**721977(*149355*)**

3.2. News Classification Models

We explore the use of a few classification algorithms to train news classification models. Specifically we train

- Logistic Regression,
- Support Vector Classification,
- XGBoost, and
- MLP Neural Network.

To deal with the challenge of having a small amount of data on short text, we use data augmentation methods, specifically a word embedding based augmentation (Wang

and Yang, 2015), approach that has been shown to work well on short text (Marivate and Sefara, 2019). We use this approach since we are not able to use other augmentation methods such as synonym based (requires developed Wordnet Synsets (Kobayashi, 2018)), language models (larger corpora needed train) and back-translation (not readily available for South African languages). We develop and present the use of both word and document embeddings (as an augmentation quality check) inspired by a hierarchical approach to augmentation (Yu et al., 2019).

4. Experiments and Results

This Section presents the experiments and results. As this is still work in progress, we present some avenues explored in both training classifiers and evaluating them for the task of news headline classification for Setswana and Sepedi.

4.1. Experimental Setup

For each classification problem, we perform 5 fold cross validation. For the bag-of-words and TFIDF vectorizers, we use a maximum token size of 20,000. For word embeddings and language embeddings we use size 50. All vectorizers were trained on the large corpora presented earlier.

4.1.1. Baseline Experiments

We run the baseline experiments with the original data using 5-fold cross validation. We show the performance (in terms of weighted F1 score) in the Figures 5 and 6. We show the baseline results as *orig*. For both the Bag-of-Words (TF) and TFIDF, the MLP performs very well comparatively to the other methods. In general the TFIDF performs better.

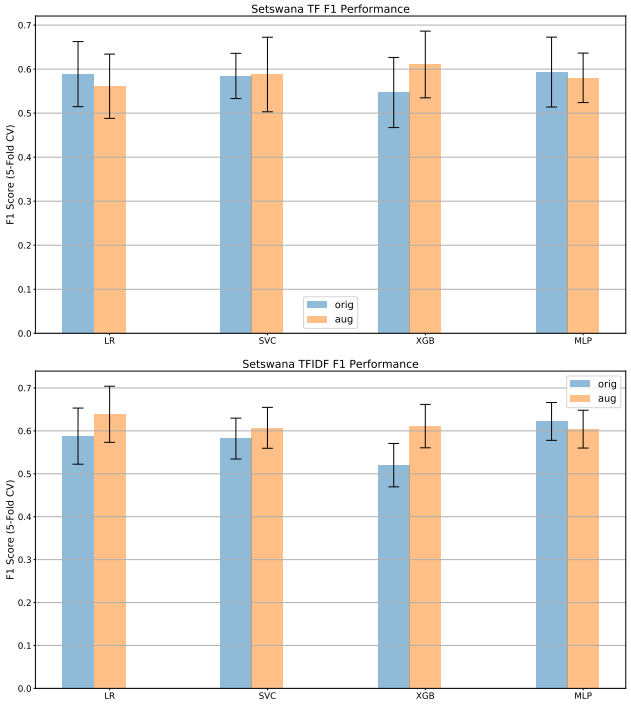

Figure 5: Baseline classification model performance for Setswana news title categorization

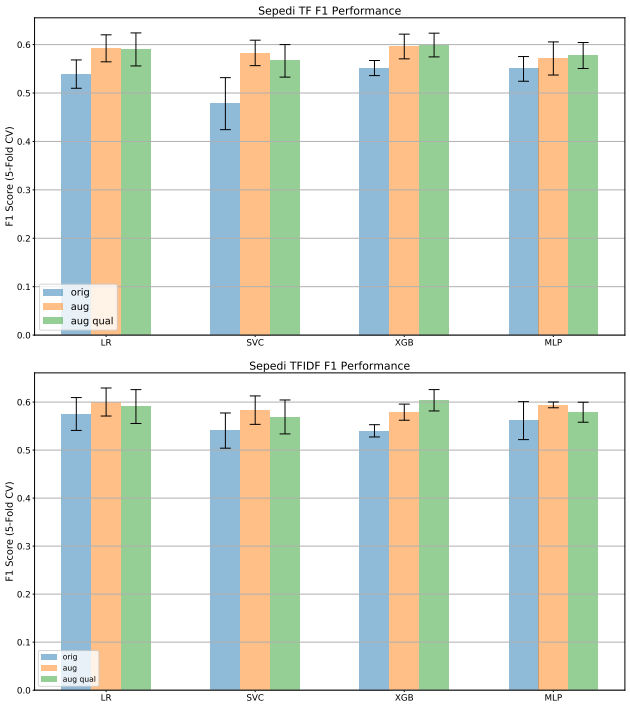

Figure 6: Baseline classification model performance for Sepedi news title categorization

4.1.2. Augmentation

We applied augmentation in different ways. First for Sepedi and Setswana word embeddings (word2vec), we use word embedding-based augmentation. We augment each dataset 20 times on the training data while the validation data is left intact so as to be comparable to the earlier baselines. We show the effect of augmentation in Figures 5 and 6 (performance labeled with *aug*).

The contextual, word2vec based, word augmentation improves the performance of most of the classifiers. If we now introduce a quality check using doc2vec (Algorithm 1) we also notice the impact on the performance for Sepedi (Figure 6 *aug qual*). We were not able to complete experiments with Setswana for the contextual augmentation with a quality check, but will continue working to better under stand the impact of such an algorithm in general. For example, it remains further work to investigate the effects of different similarity thresholds for the algorithm on the overall performance, how such an algorithm works on highly resourced languages vs low resourced languages, how we can make the algorithm efficient etc.

It also interesting to look at how performance of classifiers that were only trained with word2vec features would fair. Deep neural networks are not used in this current work and as such we did not use recurrent neural networks, but we can create sentence features from - word2vec by either using: the mean of all word vectors in a sentence, the median of all word vectors in a sentence or the concatenated power means (Rücklé et al., 2018). We show the performance of using this approach with the classifiers used for Bag of Words and TFIDF earlier in Figure 7.

The performance for this approach is slightly worse with

Algorithm 1: Contextual (Word2vec-based) augmentation algorithm with a doc2vec quality check

Input: s: a sentence, run: maximum number of attempts at augmentation
Output: $\hat{s}$ a sentence with words replaced

def `Augment(`*s,run*`)`**:**
  Let $\vec{V}$ be a vocabulary;
  for i **in** `range(`*run*`)` **:**
    $w_i \leftarrow$ randomly select a word from s;
    $\vec{w} \leftarrow$ find similar words of w_i;
    $s_0 \leftarrow$ randomly select a word from $\vec{w}$ given weights as distance;
    $\hat{s} \leftarrow$replace w_i with similar word s_0;
    $\vec{s} \leftarrow Doc2vec(s)$;
    $\vec{\hat{s}} \leftarrow Doc2vec(\hat{s})$;
    $similarity \leftarrow$ Cosine Similarity($\vec{s}$, $\vec{\hat{s}}$);
    if $similarity > threshold$ **:**
      return($\hat{s}$);

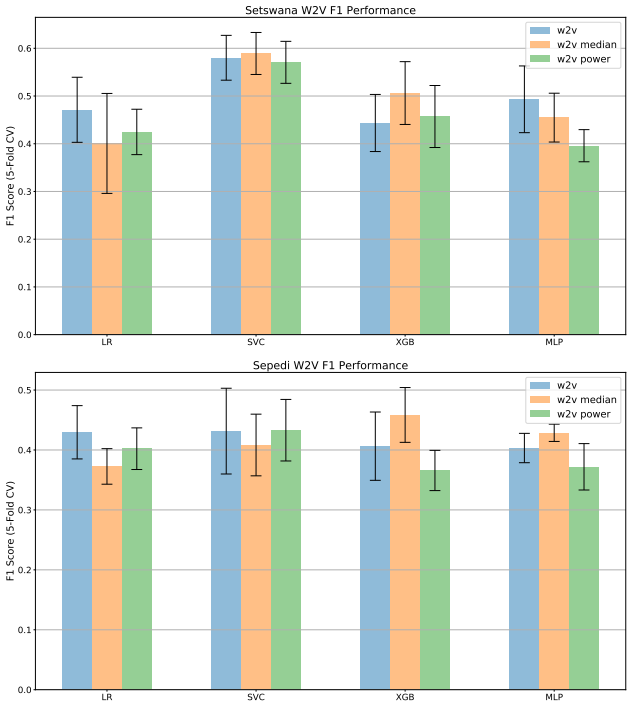

Figure 7: Word2Vec feature based performance for news headline classification

the best results for Sepedi news headline classification being with XGBoost on the augmented data. We hope to improve this performance using word2vec feature vectors using recurrent neural networks but currently are of the view that increasing the corpora sizes and the diversity of corpora for the pre-trained word embeddings may yield even better results.

Finally, we show the confusion matrix of the best model in Sepedi on a test set in Figure 8. The classifier categorizes *General News*, *Politics* and *Legal* news headlines best. For others there is more error. A larger news headline dataset is required and classification performance will also

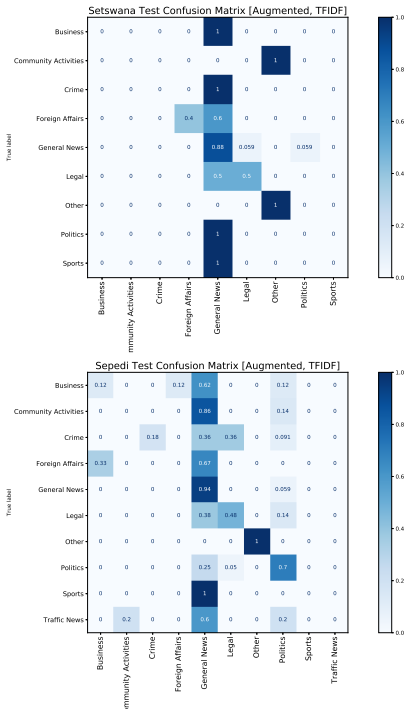

Figure 8: Confusion Matrix of News headline classification models

need to be compared to models trained on full news data (with the article body). For the Setswana classifiers, the confusion matrix shows that the data skew results in models that mostly can categorize between categories *General News* and *Other*. We need to look at re-sampling techniques to improve this performance as well as increasing the initial dataset size.

5. Conclusion and Future Work

This work introduced the collection and annotation of Setswana and Sepedi news headline data. It remains a challenge that in South Africa, 9 of the 11 official languages have little data such as this that is available to researchers in order to build downstream models that can be used in different applications. Through this work we hope to provide an example of what may be possible even when we have a limited annotated dataset. We exploit the availability of other free text data in Setswana and Sepedi in order to build pre-trained vectorizers for the languages (which are released as part of this work) and then train classification models for news categories.

It remains future work to collect more local language news headlines and text to train more models. We have identified other government news sources that can be used. On training embedding models with the data we have collected, further studies are needed to look at how augmentation using the embedding models improve the quality of augmentation.

6. Bibliographical References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword informa-

tion. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bureau of Circulations, A. (2019). Newspaper circulation statistics for the period January-March 2019 (ABC Q1 2019).

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 113–123.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 452–457.

Lewis, D. D. (1997). Reuters-21578 text categorization collection data set.

Marivate, V. and Sefara, T. (2019). Improving short text classification through global augmentation methods. *arXiv preprint arXiv:1907.03752*.

Marivate, V. and Sefara, T. (2020a). African embeddings [nlp]. `https://doi.org/10.5281/zenodo.3668481`, February.

Marivate, V. and Sefara, T. (2020b). South African news data dataset. `https://doi.org/10.5281/zenodo.3668489`.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Nettle, D. (1998). Explaining global patterns of language diversity. *Journal of anthropological archaeology*, 17(4):354–374.

Rücklé, A., Eger, S., Peyrard, M., and Gurevych, I. (2018). Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.

Sandhaus, E. (2008). The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Silfverberg, M., Wiemerslage, A., Liu, L., and Mao, L. J. (2017). Data augmentation for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 90–99.

Strassel, S. and Tracey, J. (2016). Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280.

Wang, W. Y. and Yang, D. (2015). That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563.

Wei, J. and Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6383–6389.

Yu, S., Yang, J., Liu, D., Li, R., Zhang, Y., and Zhao, S. (2019). Hierarchical data augmentation and the application in text classification. *IEEE Access*, 7:185476–185485.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.