Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5624–5630
Marseille, 11–16 May 2020

Language Proficiency Scoring

Cristina Arhiliuc, Jelena Mitrović, Michael Granitzer
Faculty of Computer Science and Mathematics
University of Passau, Germany
arhili01@ads.uni-passau.de, jelena.mitrovic@uni-passau.de, michael.granitzer@uni-passau.de

Abstract

The Common European Framework of Reference (CEFR) provides generic guidelines for evaluation of language proficiency. Nevertheless, for automated proficiency classification systems, different approaches for different languages are proposed. Our paper evaluates and extends the results of an approach to Automatic Essay Scoring proposed as a part of the REPROLANG 2020 challenge. We provide a comparison between our results and the ones from the original paper, and we also include experiments on a new corpus for the English language. Our results are lower than the expected when using the same approach, and the system does not scale well with the added English corpus.

Keywords: AES, Automatisation, Language Proficiency

1. Introduction

In the world of globalization and internationalization being multilingual allows for more business opportunities. This drives more individuals to learn additional languages, which in turn increases the number of language exams such as TOEFL and IELTS for English, TCF, DELF and DALF for French, telc, TestDaF, and Goethe-Institut for German, taken a few times every year.

The Common European Framework of Reference (CEFR) offers a generalized scoring system of language proficiency of learners that consists of 6 levels independent of the language: A1, A2, B1, B2, C1 and C2. Automated Essay Scoring (AES) represents the task of automatically assessing texts written by learners using natural language processing tools. The verification and validation of a new AES approach are part of the REPROLANG 2020 challenge[1] along with many other research topics in the area of natural language processing.

The goal of our work is to reproduce the results published in the original, candidate paper (Vajjala and Rama, 2018), that explores the possibility of a multilingual approach of classifying texts and to extend their approach with a new corpus. A multilingual model represents a model trained on multiple languages and capable of classifying texts in multiple languages. In our paper, we discuss several issues:

- Would building a multilingual model instead of a monolingual one have a great impact on the prediction metrics?
- Which features could improve the prediction metrics for multilingual models? What is their impact on the monolingual model?
- What are the limitations of the current model and how can it be improved?

The remainder of our paper is organized as follows. Section 2. gives a short overview of the State of the Art research on AES approaches. A short description of the used corpora is presented in section 3., followed by the methodology applied in this paper in section 4. Section 5. shows the results of reproducing the original paper's experiments. Furthermore, section 6. describes the cross-lingual experiments. Additionally, the data-set is augmented and experimented with in section 7. Lastly, we give conclusions relevant to our research in section 9.

2. State Of The Art

Common approaches to building AES systems are based on monolingual evaluation (Alikaniotis et al., 2016; Yannakoudakis et al., 2011). Monolingual evaluation focuses on the language particularities and yields good results. However, new approaches that construct and evaluate models on multiple languages are emerging, as presented in the original paper (Vajjala and Rama, 2018) we base our work upon. The authors investigate the possibility of building a universal CEFR classifier and analyze three categories of classification:

- Monolingual classification: Training and evaluating classifiers on texts written in the same language;
- Multilingual classification: Training and evaluating classifiers on texts written in multiple languages;
- Cross-lingual classification: Training classifiers on one language and evaluating them on other languages.

Their experiments were conducted on a multilingual corpus called MERLIN (Boyd et al., 2014), especially on 3 languages: German, Czech and Italian. Each text is enriched with metadata, such as information about the author, information about the text and CEFR levels of rating criteria.

The original corpus was transformed into text files, which contain only the texts without any metadata. The names of these files contain information about the language and CEFR level. In order to tag parts of speech from the corpus texts the authors used the UDPipe parser (Straka et al., 2016) with universal dependencies treebanks (Nivre et al., 2016). With these tools parsed files were created in separate directories.

[1] REPROLANG 2020, https://www.clarin.eu/event/2020/reprolang-2020, last accessed on July 21, 2019

The authors further emphasized several AES specific features to evaluate a text independent of the language:

- Word and POS n-grams, which are common in AES classifiers (Yannakoudakis et al., 2011);
- Embeddings of task-specific words and characters trained through a softmax layer. The authors pointed out that their paper is the first to explore character embeddings as a cross-linguistic feature for AES classifiers;
- Dependency n-grams where each unigram consists of 3 elements: The dependency relation, the POS tag of the dependent and the POS tag of the head. The authors pointed out that these features were not used in previous work on AES systems.
- Linguistic features such as:
  - Document length: The number of words in a text;
  - Lexical richness features: Lexical density, lexical variation and lexical diversity features;
  - Error features: These are obtained by using LanguageTool[2] for spelling and grammar checking. These features were only collected for German and Italian.

Logistic regression, Random Forest, Multi-layer Perceptron and SVM are compared on experiments with non-embedding features. For the embedded features, neural network models are trained specifically for that task. They use categorical cross-entropy loss and Adadelta algorithm to train the algorithm. For classification with word embeddings, they used a softmax layer.
They considered 2 different categories of features when experimenting with classifiers:

- Non-embedding features - used for Logistic Regression, Random Forest, Multi-layer Perceptron and SVM implemented using scikit-learn[3];
- Embedding features - neural network models are implemented using Keras[4] with TensorFlow[5] as backend.

The results of their experiments were measured using a weighted F1 score. The purpose is to compute the weighted average of the F1 score taking class distribution into account.

3. Datasets Presentation

In the original paper (Vajjala and Rama, 2018), the MERLIN dataset (Boyd et al., 2014) was used. It contains 2,286 texts which were taken from written examinations of acknowledged test institutions. This dataset contains texts in 3 languages: Czech, German and Italian. Every text is overall graded according to CEFR.
For the purposes of preprocessing the data, the text files from levels in which there were less than 10 instances were removed from the dataset. Furthermore, unlabeled files were also removed. The final version of the corpora consisted of 2267 texts, the distribution of which is shown in Table 1.
International Corpus Network of Asian Learners of English (ICNALE)[6] offers freely available text corpora graded according to the CEFR levels. They contain several collections of different kinds of texts and speech collected from learners of the English language in 10 Asian countries and regions, as well as from native speakers. For the purpose of this project, the ICNALE Written Essays module (Ishikawa, 2013), containing 5600 essays (200-300 words long) about two topics, is used. For the experiments, only 5200 essays are used and 400 were removed due to missing labels.
The distribution of labels in the new dataset that we experimented on is shown in the last column of table 1. The English corpus contains files labeled as A2, B1 and B2 only. The issue of not having texts of all labels is also present in the MERLIN dataset.

CEFR level	CZ	DE	IT	EN
A1	0	57	29	0
A2	188	306	381	960
B1	165	331	394	3776
B2	81	293	0	464
C1	0	42	0	0
Total	434	1029	804	5200

Table 1: Distribution of labels in corpora

4. Methodology

The authors of the original paper approached the topic of AES systems differently from how it was done in previous work in that:

- They use the CEFR system to study the AES systems;
- They explore the possibility of a Universal AES, given that the CEFR guidelines are not language specific. They call it the Universal CEFR classifier;
- They are exploring cross-lingual AES.

The goal of our research is to verify the results published in (Vajjala and Rama, 2018) and to experiment with an additional language. Therefore, three tasks are required:

1. The mentioned experiments will be reproduced, monitored and documented and the results will be compared using the provided code[7];

[2] LanguageTool, https://languagetool.org/, last accessed on July 21, 2019

[3] scikit-learn, https://scikit-learn.org/stable/, last accessed on July 21, 2019

[4] Keras, https://keras.io/, last accessed on July 21, 2019

[5] TensorFlow, https://www.tensorflow.org/, last accessed on July 21, 2019

[6] ICNALE: The International Corpus Network of Asian Learners of English, http://language.sakura.ne.jp/icnale/, last accessed on July 21, 2019

[7] GitHub repository, https://github.com/nishkalavallabhi/UniversalCEFRScoring, last accessed on July 21, 2019

Features	DE	IT	CZ	Avg. Dev.
Baseline	0.477 (-0.020)^{RF}	0.573 (-0.005)^{LR}	0.613 (+0.026)^{LR}	0.017
Word n-grams(1)	0.589 **(-0.077)**^{RF}	0.799 (-0.028)^{RF}	0.727 (+0.006)^{RF}	0.037
POS n-grams(2)	**0.658** (-0.005)^{RF}	0.801 (-0.024)^{RF}	0.678(-0.021)RF	0.016
Dep. n-grams(3)	0.637 (-0.026)^{RF}	0.800 (-0.006)^{RF}	0.706 (+0.002)^{RF}	0.011
Domain features	0.520 (-0.013) ^{LR}	0.654 (+0.001)^{LR}	0.629 **(-0.034)**^{RF}	0.016
(1)+Domain	0.644 (-0.042)^{RF}	0.793**(-0.044)**^{RF}	0.720 (-0.014)^{RF}	0.033
(2)+Domain	0.646 (-0.040)^{RF}	0.796 (-0.020)^{RF}	0.687 (-0.022)^{RF}	0.027
(3)+Domain	0.639 (-0.043)^{RF}	0.784 (-0.022)^{RF}	**0.730** (+0.018)^{RF}	0.027
Word embeddings	0.633 (-0.013)	**0.804** (+0.010)	0.653(+0.028)	0.017
Avg. Dev.	0.028	0.016	0.017	

Table 2: Weighted F1 scores for monolingual Classification compared to the results from (Vajjala and Rama, 2018) (in parenthesis, $value_{reproduced} - value_{original}$).

Features	Lang (-)	Lang (+)	Avg. Dev.
Baseline	0.426 (-0.002)^{LR}	-	0.002
Word n-grams	0.605 **(-0.116)**^{RF}	0.607 **(-0.112)**^{RF}	0.114
POS n-grams	0.680 (-0.046)^{RF}	0.680 (-0.044)^{RF}	0.045
Dep. n-grams	0.650 (-0.053)^{RF}	0.652 (-0.041)^{RF}	0.047
Domain features	0.433 (-0.016)^{LR}	0.447 (-0.024)^{LR}	0.020
Word embeddings	**0.683** (-0.010)	**0.681** (-0.008)	0.009
Avg. Dev.	0.040	0.038	

Table 3: Weighted F1 scores for multilingual classification with models trained on combined datasets compared to the results from (Vajjala and Rama, 2018) (in parenthesis, $value_{reproduced} - value_{original}$).

2. An English corpus from (Ishikawa, 2013) will be added to the dataset, the experiments will be executed again and the results will be reported in this paper.

3. Experiments with cross-lingual classifiers using inter-family and intra-family languages will also be performed;

Throughout this paper and in the provided tables, the notations RF, LinSVC and LR are indicating the used classifiers: Random Forest, Linear Support Vector Classifier or Logistic Regression respectively.

Our tests, except for the word embeddings, were done on a machine with the following configuration: processor Intel(R) Core(TM) i7-7700HQ CPU @ 2.8 GHZ 3.8 GHZ; the RAM of the machine is 16 GB (15.9 GB usable) and the operating system is a 64-bit Windows 10 Home Edition, x64-based processor. For the word embeddings, an Nvidia Tesla K80 GPU was used and 251 GiB made available given that the results on CPU for word embeddings were much lower than the ones in the original paper. We have no information about the hardware used for the execution of the experiments whose results are reported in the original paper, therefore we could not assess if the differences in hardware had anything to do with the differences in the results we have achieved, compared to the results of the original paper.

The programming language used for the experiments mentioned in this paper is Python 3.7. We have no information about the version of Python and the versions of the libraries used for the experiments from (Vajjala and Rama, 2018). The environment file for our execution is in the GitLab repository [8]

5. Analysis

For the purpose of reproducing the experiment depicted in the original paper, we use the environment described in section 4. Our main goal in this regard is to check the validity of the results presented in their paper and to explore possibilities for improvement.

Throughout this section, the average deviation is defined as follows:

$$Average\ Deviation = \frac{\sum |value_{reproduced} - value_{original}|}{n}$$

where n is the number of values in the column or in the row. We are going to compare the original paper's results with our own for monolingual, multilingual and cross-lingual classification with German as the training language.

Table 2 presents the results that we have obtained during monolingual classification. The results that we get are different from the original results. The biggest difference we notice is in the classification based on word n-grams

[8] https://gitlab.com/Cristinutaa1/language_proficiency_scoring

Features	**Test: IT**	**Test: CZ**	**Avg. Dev.**
Baseline	0.553 (=)^{LR}	0.48 (=)^{LR}	0.000
POS n-grams	**0.752** (-0.006)^{RF}	**0.679 (+0.030)**^{RF}	0.018
Dep. n-grams	0.60 **(-0.023)**^{RF}	0.66 (-0.012)^{RF}	0.017
Domain features	0.62 (-0.001)^{LR}	0.46 (-0.009)^{RF}	0.005
Avg. Dev.	0.007	0.017	

Table 4: Weighted F1 scores for cross-lingual classification model trained on German compared to the results from (Vajjala and Rama, 2018) (in parenthesis, $value_{reproduced} - value_{original}$).

for German language (-0.077) and the biggest difference in terms of better results is for Czech language (+0.026). The word n-grams based classification seems to give the results that are the furthest from the original paper, with an average deviation of 0.037. As in the original results, the n-grams seem to perform better than syntactic features. Although, here, they seem to be better than or to have close results with the combination between n-grams and domain. In our experiments, this combination doesn't seem to have a significant improvement on the results (e.g. German with dependency n-grams) in some situations and impacts them negatively in some other scenarios (e.g. Italian with dependency n-grams). It can also be noticed that the model performs worse for German, which could be explained by the greater number of classes. Table 3 presents the results obtained during multilingual classification. We notice a significant difference between our results and the ones from the paper for word n-grams both with and without language features. We compared the results published in the paper with the ones available in the "Results" directory of the source code and they were identical. We could not find the source of this problem, especially given that the same function for treating word n-grams was used for both multilingual classification and monolingual classification and this function gives close results for Italian and Czech in monolingual classification. Our results concerning the best features for the multilingual classification differ from the original paper: word embeddings perform the best here, followed by POS n-grams, although our F1 score is lower. For all multilingual experiments, our F1 score was lower than in the original paper. The closest results we got are for the baseline and word embeddings. The multilingual model's F1 score is not much lower than the average of monolingual results for these languages, which leads to the conclusion that at least for these three languages the experiment is a success.

→ Pred	**A1**	**A2**	**B1**	**B2**	**C1**
A1	3 (-2)	26 (+2)	0	0	0
A2	9 (=)	330 (+19)	39 (-17)	3 (-2)	0
B1	2 (-1)	89 (+19)	260 (-19)	43 (-1)	0

Table 5: DE-Train:IT-Test setup with POS n-gram features compared to the results from (Vajjala and Rama, 2018) (in parenthesis, $value_{reproduced} - value_{original}$).

→ Pred	**A1**	**A2**	**B1**	**B2**	**C1**
A2	0	134 (+5)	54 (-3)	0 (-2)	0
B1	0	30 (+7)	98 (-3)	37 (-4)	0
B2	0	2 (-3)	24 (-1)	55 (+4)	0

Table 6: DE-Train:CZ-Test setup with dependency features compared to the results from (Vajjala and Rama, 2018) (in parenthesis, $value_{reproduced} - value_{original}$).

Tables 4, 5, 6 present the results of cross-lingual classification. We can notice in table 4 that the results of our experiments are close to the ones from the original paper. The only notable difference could be seen in tables 5 and 6 for the predicted A2, where in our environment the classifiers seem to predict more A2 for texts with the true labels A1 and B1. Nevertheless, our experiments showed lower misclassfications of A2 as B2 and of B2 as A2. The quality of the classification is good, which suggests that the given features are indeed cross-lingual (are valid and similar for multiple languages).

Overall, our experiments have shown worse results compared to the results in the original paper. We assume that one possible reason for this is that the results published in the original paper were obtained by repeating the experiments multiple times and choosing the best results, having a different hardware configuration or because of the random factors influencing the models. Our results were obtained only as a result of one execution. Nevertheless, the results of the experiment for these 3 languages, multilingual and cross-lingual with German as the training language, proved to be successful.

6. Cross-lingual extension

Given that the authors of the original paper have published the results only for cross-lingual classification with German as training language, we decided to extend their experiments for the other languages to check wether the results of the system would be similar.

Features	Test: DE	Test: CZ
Baseline	**0.711**$^{\textbf{LinSVC}}$	**0.770**$^{\textbf{LR}}$
POS n-grams	0.508$^{\text{RF}}$	0.657$^{\text{RF}}$
Dep. n-grams	0.549$^{\text{LinSVC}}$	0.602$^{\text{LinSVC}}$
Domain features	0.706 $^{\text{LinSVC}}$	0.756$^{\text{RF}}$

Table 7: Weighted F1 scores for cross-lingual classification model trained on Italian.

Features	Test: DE	Test: IT
Baseline	**0.528**$^{\textbf{RF}}$	0.697$^{\text{LR}}$
POS n-grams	0.444$^{\text{LR}}$	0.587$^{\text{RF}}$
Dep. n-grams	0.363$^{\text{LR}}$	0.531$^{\text{RF}}$
Domain features	0.478 $^{\text{LR}}$	**0.796**$^{\textbf{LinSVC}}$

Table 8: Weighted F1 scores for cross-lingual classification model trained on Czech.

The authors explain that they used the German texts corpus for training because that was the only corpus containing samples of all the labels. In order to make the cross-language validation correct, we made sure that the prediction was made only on the segment of the data that has the same labels as the training language.

As presented in tables 7 and 8, we noticed that the results of using languages other than German (i.e. Italian and Czech) for training are quite different. Baseline and domain features seem to perform better in these two cases, which suggests that the length of the text has a big impact on the result of the classification. In this case, the model will work well when the text requirements specify different lengths for different levels, but may fail when the length of the text is similar, but the content is of different level.

In tables 9 and 10 where the test data had more than two labels, we can notice the same tendency of the classifier to predict the label A2 more often for the texts with the true labels A1 and B1. If we consider the F1 score per label, A2 has the greatest value. When the number of examples per label is balanced as in table 10, the model seems to underrate the example.

→ Pred	A1	A2	B1
A1	2	55	0
A2	0	227	79
B1	0	47	284

Table 9: IT-Train:DE-Test setup with baseline

→ Pred	A2	B1	B2
A2	283	12	3
B1	186	106	39
B2	23	146	124

Table 10: CZ-Train:DE-Test setup with baseline

7. Experiments With Augmented Dataset

One of the aims of the paper was to check the validity of the approach proposed in (Vajjala and Rama, 2018) for other languages and to investigate if building classifiers on languages belonging to the same family group improves the results for cross-lingual classification. For this reason, a new corpus for the English language is added to the corpora list on which experiments are executed. The files are renamed and parsed using the English treebank[9] and the same UDPipe tool in order to have the same structure and information as the ones for the other languages.

Features	English
Baseline	0.333$^{\text{LinSVC}}$
Word n-grams(1)	0.617$^{\text{RF}}$
POS n-grams (2)	0.615$^{\text{RF}}$
Dep. n-grams(3)	0.616$^{\text{RF}}$
Domain features	0.335$^{\text{LinSVC}}$
(1) + domain	0.629$^{\text{RF}}$
(2) + domain	0.620$^{\text{RF}}$
(3) + domain	0.620$^{\text{RF}}$
Word embeddings	**0.640**

Table 11: Weighted F1 scores for English monolingual classification

Table 11 shows results for the monolingual classification based on the English data set. We can see that performance for features such as Word n-grams, POS n-grams and dependency n-grams is doubled compared to the baseline. Additionally, the variation among the mentioned features is minor because of the size of the new data set that is at least 4 times bigger than the original data sets. The baseline did not perform well, which was expected given that it is based on the lengths of the documents. According to the table 12, for the English corpus, the variation of text lengths for different labels is insignificant and therefore a bad criterion for classification. Adding domain features is not improving significantly the result because it is mainly dependent on the length of the text.

CEFR level	CZ	DE	IT	EN
A1	-	32.23	39.86	-
A2	93.68	56.89	69.04	214.28
B1	169.81	112.48	145.61	224.54
B2	205.91	187.96	-	232.92
C1	-	220.95	-	-

Table 12: Average document length per CEFR level

Table 13 presents the result of multilingual classification that was extended with English language. We compared

[9] UDPipe model for English, https://github.com/jwijffels/udpipe.models.ud.2.0/tree/master /inst/udpipe-ud-2.0-170801, last accessed on July 21, 2019

new results with obtained results from section 5. both with and without language features. The difference between the current value and the one obtained in section 5. is indicated in the parentheses.

We observed that adding the English corpus had a negative effect on all F1 scores. However, we want to emphasize that dependency n-grams features and POS n-grams features have the smallest average deviation (0.029 and 0.046) given that the pattern of sentence structure is similar for the majority of the European languages. Simultaneously, baseline and domain features have the biggest average deviations (0.118 and 0.098) which could be explained by the fact that the majority of data samples come from English corpus and English text files classification performs poorly for baseline and domain features.

Table 14 presents F1 scores for cross-lingual classification model trained on English corpus. We performed this step to check if a language from the same family would improve performance. The authors of the original paper mentioned that word n-grams and word embeddings are not suitable for cross-language classification. Therefore, the considered features are: baseline, domain features, POS n-grams and dependency n-grams. We noticed that results for a language from the same family (tested on German) have lower F1 scores especially for dependency n-grams and domain features compared to inter-family (tested on Italian and Czech).

Features	Test: DE	Test: IT	Test: CZ
Baseline	0.272^{LR}	$\mathbf{0.726^{LR}}$	0.536^{LR}
POS n-grams	0.431^{RF}	$\mathbf{0.821^{RF}}$	0.570^{RF}
Dep. n-grams	0.299^{LinSVC}	$\mathbf{0.580^{LinSVC}}$	0.351^{RF}
Domain features	0.289^{LR}	$\mathbf{0.363^{LR}}$	0.242^{LR}

Table 14: Weighted F1 scores for cross-lingual classification model trained on English

Table 15 shows a confusion matrix based on POS n-grams features. Moreover, the results from multilingual classification have better scores than the ones from cross-lingual classification.

→ **Pred**	**A2**	**B1**	**B2**
A2	226	40	0
B1	112	219	0
B2	4	289	0

Table 15: EN-Train:DE-Test setup with POS n-gram features

However, results of testing on Italian demonstrate the best performance among other languages and POS n-grams show great performance with an F1 score of 0.821.

The confusion matrix in table 16 shows the misclassification only for adjoining levels of proficiency. Additionally, results of testing on Italian also demonstrate better performance for baseline and POS n-grams in comparison to multilingual classification.

→ **Pred**	**A2**	**B1**	**B2**
A2	328	53	0
B1	85	309	0

Table 16: EN-Train:IT-Test setup with POS n-gram features

Now we consider table 17 consisting of F1 scores for cross-lingual classification tested on English texts. As in the previous description, we did not consider all features but only baseline, POS n-grams, dependency n-grams and domain features. We observed that baseline and domain features have the smallest values (0.075 and 0.107) in training on German. This means that the model performs poorly. The reason for this is that as mentioned before, the lengths of essays in English and in German vary dramatically according to table 12. Results of training on Italian (Table 17) show great performance as well as results of testing on Italian (Table 14). F1 scores of using dependency n-grams and domain features have better effectiveness (20% and 45%).

Features	Train: DE	Train: IT	Train: CZ
Baseline	0.075^{RF}	$\mathbf{0.707^{LinSVC}}$	0.400^{RF}
POS n-grams	0.362^{RF}	$\mathbf{0.716^{RF}}$	0.567^{RF}
Dep. n-grams	0.449^{LR}	$\mathbf{0.718^{RF}}$	0.619^{RF}
Domain features	0.107^{RF}	$\mathbf{0.708^{RF}}$	0.614^{RF}

Table 17: Weighted F1 scores for cross-lingual classification model tested on English

Furthermore, we discovered an interesting case that the efficiency of domain features is different for testing on Czech (Table 14) and training on Czech (Table 17). The performance of domain features of training on Czech is almost 3 times better that of testing on Czech.

The confusion matrices 18 and 19 demonstrate that cross-lingual classification on texts with the true label B1 between English and Czech performs poorly: especially for classifiers trained on English that have an accuracy of at most 25%.

→ **Pred**	**A2**	**B1**	**B2**
A2	5	183	0
B1	0	164	1
B2	0	79	2

Table 18: EN-Train:CZ-Test setup with domain features

→ **Pred**	**A2**	**B1**	**B2**
A2	12	935	13
B1	22	3630	124
B2	0	432	32

Table 19: CZ-Train:EN-Test setup with domain features

Features	Lang (-)	Lang (+)	Avg. Dev.
Baseline	0.308 **(-0.118)**^{LR}	-	0.118
Word n-grams	0.563 (-0.042)^{RF}	0.559 (-0.048)^{RF}	0.045
POS n-grams	**0.634** (-0.046)^{RF}	**0.634** (-0.046)^{RF}	0.046
Dep. n-grams	0.623 (-0.027)^{RF}	0.620 (-0.032)^{RF}	0.029
Domain features	0.318 (-0.115)^{LR}	0.365 **(-0.082)**^{LR}	0.098
Word embeddings	0.596 (-0.087)	0.591 (-0.090)	0.088
Avg. Dev.	0.071	0.056	

Table 13: Weighted F1 scores for multilingual classification with models trained on Italian, Czech, German and English corpora, compared to the ones trained on Italian, Czech and German corpora

8. Discussions

By examining and reproducing the results of the original paper by Vajjala and Rama, we have arrived to the same conclusion - it is worth further investigating the possibility of multilingual classifiers, because in this case the results were not significantly worse than for monolingual. However, it would be interesting to see if by adding more languages the results don't become worse. Do we reach a plateau, or would the results keep decreasing ? It is also worth checking what happens when adding a non-European language. These questions are all related to how much the model could be extended.

Another idea that is worth looking into is transforming the problem into a regression problem, which could lead to a better view of the examples that are between classes. This could be used directly in practice in a semi-automated system where the examples that are at the border can be manually examined. Also, it can lead to some ideas of improvement of the current system.

The model could also be improved by exploring other relevant features and the possible enhancement they could give compared to the generic model. Certain new features related to the semantic and syntactic analysis of the texts - correctness of the sentence structure and sentence meaning score - could prove advantageous for further approaches.

In terms of meaning, the correlation between a text and a task is not considered, thus even if a text is not related to the demanded task, it can be highly graded. This is also something that is worth exploring.

9. Conclusion

Following the execution of the same experiments as the ones presented in (Vajjala and Rama, 2018), our results showed a significant difference to the published ones, especially on the multilingual models: for the word n-grams we have an average deviation between our results and the ones from the original paper of 0.114, 0.047 for the dependence n-grams and 0.045 for the POS n-grams. We were not able to find an explanation for this difference in the results.

Furthermore, our paper proved that the presented approach does not scale well when English is added. This could have been caused by the different properties of the English corpus: text length, lexical diversity and sentence structure.

Our experiments show that for these corpora of English and German there are no better results in intra-family classification, as we would have expected. The results have nevertheless proven a good correlation between English and Italian. The cross-validation models depend too much on the combination of languages, which makes them difficult to generalize.

10. Bibliographical References

Alikaniotis, D., Yannakoudakis, H., and Rei, M. (2016). Automatic text scoring using neural networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).*

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Stindlová, B., and Vettori, C. (2014). The merlin corpus: Learner language and the cefr. In *LREC*, pages 1281–1288.

Ishikawa, S. (2013). The icnale and sophisticated contrastive interlanguage analysis of asian learners of english. *Learner corpus studies in Asia and the world*, 1:91–118.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-u files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4290–4297, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Vajjala, S. and Rama, T. (2018). Experiments with universal CEFR classification. *CoRR*, abs/1804.06636.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 180–189, Stroudsburg, PA, USA. Association for Computational Linguistics.