Predicting Educational Achievement Using Linear Models

Çağrı Çöltekin
Department of Linguistics
University of Tübingen
ccoltekin@sfs.uni-tuebingen.de

Abstract

This paper describes our participation in GermEval 2020 task 1 on the classification and regression of cognitive and motivational style from text, which includes two subtasks. The first subtask is about predicting ranking of students based on a number of academic achievement scores, and the second subtask is about predicting the categorical outcomes of a psychometric test. The systems used in this study are based on simple linear models trained only on the task data for both systems. Despite their simplicity, our systems obtained the first place with a Pearson correlation coefficient of 0.370 in subtask 1, and the third place with a macro-averaged F1 score of 0.678 in subtask 2 (20-way classification). Besides describing the systems, we report additional analyses and results, and discuss possible improvements and ethical considerations regarding the task at hand.

1 Introduction

Language use reflects many aspects of a speaker's (or author's) psychology. As a result, linguistic output of a person can be used for detecting certain aspects of his/her psychological state (see Johannßen and Biemann, 2018, for a recent review). This has been demonstrated successfully on a range of tasks including predicting basic personal features like gender or age (Barbieri, 2008; Peersman et al., 2011; Burger et al., 2011; Nguyen et al., 2014); predicting personality traits (Luyckx and Daelemans, 2008; Celli et al., 2013; Plank and Hovy, 2015); predicting sentiment towards the topic of a text (Pang et al., 2008), and predicting mental health (Ramirez-Esparza et al., 2008; Coppersmith et al., 2014; Pirina and Çöltekin, 2018).

The present shared task (Johannßen et al., 2020) is about predicting scores and psychometrics used for assessing academic achievement. The task consist of two subtasks. Subtask 1 requires predicting a ranking of students based on the sum of high school grades and IQ test scores. The source for the predictions consists of short texts obtained during a psychometric test. In subtask 2, the aim is to predict the outcome of another psychometric test from similar short texts. Both subtasks have attracted interest from researchers. For example, Pennebaker et al. (2014) study the correlation between word choices in college application essays and later academic success, and Johannßen et al. (2019) present methods for automatically annotating a version of the subtask 2 data set (with different class labels).

Although the task as formulated in the present form may not be directly applicable, automatic methods have been put in real-world use for similar purposes, such as automatic essay scoring (Dikli, 2006; Ke and Ng, 2019, see, e.g.,). As a result, the methods developed here can *potentially* be used in real-world educational assessment tasks. The potential applicability on an area with high societal impact, combined with known racial or socio-economic biases in some of the scores (e.g., IQ scores) used in the present task, raises a number of ethical concerns. A brief discussion of ethical concerns are presented in Section 5, and the task organizers review past and present use and misuse of such metrics (Johannßen et al., 2020).

The systems used in this study are simple linear models for both subtasks. For subtask 1, we use linear regression to predict the academic success indicators. For subtask 2, we use support vector machine (SVM) classifiers. For both systems, we use word and character n-gram features extracted from the shared task data provided by the organizers, without use of any external resources or linguistic processing. We mainly focus on the regression task (subtask 1) where we report additional results with a few alternative settings.

2 The Task and the Data

This section briefly describes the task and the data. A more detailed description can be found in Johannßen et al. (2020).

Subtask 1 The first subtask requires predicting an (artificial) ranking of students from their linguistic output. For each student there are 30 short texts which are answers given to questions about 15 different images during a psychometric test, *motive index* (MIX, Scheffer and Kuhl, 2006). The expected outcome is a ranking of the students obtained by summing five different scores, namely, high school German, English and math grades and scores of language IQ and logic IQ. All scores were z-normalized by the organizers.[1] The distributions of all scores show some skew. In particular, the IQ scores have a negative skew, while the high school grades have a positive skew.

The data is obtained from the application process of NORDAKADEMIE, a private university of applied sciences in Germany. The data set contains texts from approximately 2600 students (approximately 78 000 answers) which was split as training, development and test sets each containing 80 %, 10 % and 10 % of the data respectively.

Subtask 2 The second subtask is classifying short texts collected during another psychometric test, *operant motive test* (OMT, Kuhl and Scheffer, 1999). The texts are labeled using a two-dimensional scheme with dimensions *motive* and *level*. There are five motive classes in the data: power (`M`), affiliation (`A`), achievement (`L`), freedom (`F`), and no motive identified (`0`). The level variable takes values from `0` to `5`. Although levels are indicated with numeric values, the underlying descriptions make the levels difficult to place on a scale. Hence, treating them as discrete classes is reasonable. For descriptions of the data, including the descriptions of labels the reader is referred to Johannßen et al. (2020) and Kuhl (2013).

Subtask 2 data consists of answers of over 14 600 volunteers taking the OMT, and answering questions about 15 images. Neither the images and questions associated with short answers, nor (anonymous) IDs of the authors are provided in the shared task data. The data was annotated by the researchers with the motive and level classes described above. In total, the data set contains 209 000 short texts with same ratios for training, development and test sets.

[1] However, the variance of the language IQ in the training and development sets is very low (0.000 013), indicating a possible oversight during normalization.

3 System and Experiment Description

For both tasks, our participation is based on simple linear models. Contrary to our expectations, a number of neural classification and regression architectures did not perform as well during our preliminary experiments.[2] As a result, we only present results from linear models in this paper. We do not use any external data, including pre-trained embeddings of any sort, and we use the same features for both regression and classification tasks. The systems are implemented mainly with the scikit-learn library (Pedregosa et al., 2011). The source code used in this study is available at `https://github.com/coltekin/germeval2020task1`.

Features For both subtasks, we use sparse character and word (token) n-gram features. Given a text, we count all character n-grams of order one to C and word n-grams of order one to W. Both C and W are treated as hyperparameters during system tuning. We combine all features in a flat manner, and scale the features with tf-idf. Tokenization is performed with a regular expression which considers any continuous non-space character sequence as a token. Except case normalization, which is treated as a hyperparameter during tuning, no other text processing or filtering is performed.

Subtask 1: the regression model Our submissions are based on ridge regression (least-squares regression with L2 regularization), using only the textual features described above.

Besides the different feature sets, we experiment with two different types of models. In the 'factored-target' model we train a separate regression model for each score, then sum the resulting scores to obtain the final ranking. In the 'single-target' model, we predict the sum of the scores directly. Since there are 30 text samples per student, we combine each text from the same person, and

[2] This, however, should not be taken as a negative result. The reason for abandoning the neural models in this study is rather practical (mainly lack of time to tune them). Even without use of external data (e.g., embeddings), the neural architectures offer a number of advantages. For example, joint predictors with shared weights and pre-training the models in other (related) tasks are likely to be useful for both subtasks.

use this combined document as a single text instance for regression. We also experimented with training the model using each individual text as a separate training instance. We only report results with combined texts, since models trained on individual texts performed worse in our experiments.

Subtask 2: the classification model Our classifiers for OMT prediction task are SVM classifiers trained with one-vs-rest multi-class strategy. The features are the sparse n-gram features described above. We experimented with combined motive–level and individual classifiers. However, our final contribution is based on a hierarchical approach. We first train a classifier predicting only the motive, and along with the textual features we also use the distances from the decision boundary of the motive predictions as additional numeric scores for the level classifier. The software used for this task is the same as on our earlier work on other text classification tasks (Çöltekin and Rama, 2018; Çöltekin et al., 2018).

Our submission includes a simple adaptation method (similar to Jauhiainen et al., 2019; Wu et al., 2019). In adaptation mode, we train a base classifier and classify the test instances, then we re-train the classifier with an augmented training set containing a subset of test instances with highly confident predictions. In this study we fixed the definition of highly confident predictions as those predictions claimed by only one of the one-vs-rest classifiers, with a minimum distance of 0.1 from the decision boundary.

Hyperparameter tuning The systems for both tasks were tuned for maximum order of character and word n-grams and whether word features are case normalized or not. The hyperparameters of the regression models also include the L2 regularization strength and the SVM margin parameter 'C' is also tuned for the classification models. Finally, we also tune a scale parameter for the additional features used in the second layer of the hierarchical classifier. The range of hyperparameters considered are listed in Table 2 in Appendix.

During tuning, we perform a random search over the parameter space, and use 5-fold cross validation on combined training and development sets. The best parameter settings are decided based on the average of the target scores obtained in cross-validation folds. The scores optimized are the scores used for shared task evaluation, namely, correlation of the ranks for the regression task, and the macro-averaged F1-score for the classification task. For each system, we performed 4000 uniform random draws from the parameter space.[3]

Ensemble output For final predictions, we use an ensemble of 10 systems re-trained with the 10-best parameter settings obtained during tuning. For the regression model, we take the mean of the predictions from each of the top-10 models. Similarly, the ensemble output of the classifiers is based on the majority vote. In case of ties, the ensemble output is the prediction that include the prediction of the best model.

4 Results

Official results Our best submission in subtask 1 obtained the first rank in the competition with a Pearson correlation coefficient of 0.370 with 0.055 points difference from the next best score. Our best model for subtask 1 was the regression model predicting only the sum of the scores.

In subtask 2, our system is placed third. The macro-averaged F1-score obtained on the test set for predicting motive–level combination is 0.678. The rank stays the same for predicting the individual dimensions, with macro averaged F1-scores of 0.680 and 0.634 for motive and level predictions respectively. For the OMT classification task, our best model was the hierarchical classifier with adaptation. However, the adaptation made a rather small (probably non-significant) difference.

The predictability of individual scores Table 1 presents Pearson correlation coefficients between model predictions and the gold-standard values, the correlation between the model ranking and the overall rank, and standard deviation of the ranking errors. The row labels indicate the scores that each model predicts (high school grades on English, German and math, and logic and language IQ scores). The row labeled 'ensemble' indicates the combination of models trained on individual scores, while 'sum-only' presents scores of a model trained to predict the sum. The column labeled 'r' present the Pearson correlation coefficient of the predictions with actual scores predicted, 'rank r' presents correlations between predictions and the gold-standard ranking based on

[3]Except for the hierarchical OMT classifier, for which the random search was terminated after approximately 300 draws.

score	r	rank r	error std
English	0.336	0.361	85.25
German	0.274	0.312	89.21
Math	0.329	0.390	82.75
Logic	0.225	0.242	95.92
Language	0.321	0.321	90.57
Ensemble	0.431	0.415	83.39
Sum-only	0.434	0.421	82.88

Table 1: Development set scores of systems trained on the combined text and individual answers.

the sum of all scores, and 'error std' is standard deviation of the ranking error. All scores were obtained by averaging the outputs of models with 10-best hyperparameter settings identified during tuning. The systems were re-trained on the official training set and tested on the development set.

The predictions of the high school grades, particularly math and English grades result in higher rank correlations with the gold-standard ranks. The higher correlation with the gold standard scores does not uniformly transfer to their usefulness in predicting the ranks based on the sum of the scores. For example, although it is not the best predicted score, the high school math grade is a better predictor of the overall rank in comparison to other scores, also providing the lowest deviations from the gold-standard ranks. Combining different scores, either by predicting them separately or predicting their sum directly is clearly useful. Predicting the sum directly gives slightly better results than predicting individual scores and summing them. Another surprising observation is that, ranking based on math scores only result in the least ranking error, yielding even slightly better results than combinations of all scores.

Similar to the official submission, the best scores are obtained by the system predicting only the sum. However, for both scores, there is a big discrepancy between development set scores, and the official test set submissions. As noted before, the scale of the language IQ score in the data set is much lower than the other scores. Since its contribution to the sum is negligible if not rescaled, standardizing it improved our scores on the development set. However, it is likely to have a negative effect on the test set due to the mismatch of the way the sum is calculated by the system and the gold-standard data.

5 General Discussion

Results: the good and the bad The results presented in this paper and by the other participants in the shared task clearly show that there is a strong signal in the texts for predicting the overall ranking. Obtaining rank correlations up to $42\,\%$ from textual features extracted from 30 short answers is impressive. However, we should also keep in mind that this means rank predictions explain only about $16\,\%$ of the (linear) variation in the gold-standard ranking. Looking at it another way, the standard deviation of the difference between the predicted ranks by our top model based on rank correlation and the gold-standard ranking on the development set (of 260 instances) is 82.9. This is clearly better than the deviation expected from a random ranking (approximately 106). However, it also means that a large error is expected in most rank predictions. Assuming normal distribution of ranking errors, approximately $32\,\%$ of the predictions will be placed more than 82.9 ranks away from their gold-standard rank. The maximum rank difference for our best model is 217, assigning the rank 18 to the gold-standard rank 235. In summary, the results clearly are interesting as there is an unmistakable signal in the data, but it is yet far from being applicable even if we assume that the gold-standard ranking provided is a good way to rank, e.g., applications to a university.

The linear models, and beyond Another take-home message from the present results is perhaps not to dismiss simple (linear) models quickly. Although there are attractive properties of (deep) neural networks, simple linear models can yield comparable, or even better results in some problems. Furthermore, their simplicity and computational efficiency allows faster (and greener) experimentation and tuning of these systems. That being said, the flexibility of neural models allow easy incorporation of external information (e.g., pre-trained embeddings or pre-training on relevant tasks), and easier modeling multi-task learning and sharing weights across tasks. These aspects of the neural models, when used properly, are likely to improve the results presented in this paper. Other potential directions for improvements include incorporation of linguistic and/or explicit error features. Even though simple n-gram features used in this study performed well, it is difficult for these features to lead to certain general-

izations. For example, generalizing over (different types of) errors in the text is likely to be useful in this task, while individual instances of errors caught by character n-gram features are not necessarily enough for a generalization that would make use of similar errors during prediction time.

Ethical considerations Another interesting aspect of the present task is the ethical implications/considerations that resulted in a considerable debate in the computational linguistics community.[4] In particular, if such systems are used in practice, the fact that they may include certain biases which lead to discrimination (or used as justification of discrimination) is a serious concern. This is particularly true for some of the scores used for ranking in the present task setup. It is well known that IQ tests show racial, ethnic or sociocultural bias (Jensen, 1980; Rushton and Jensen, 2005, also see the task description paper by Johannßen et al. (2020) for a broad review of use and misue of IQ scores in different countries by public and private organizations). Similarly, the high-school grades are not bias-free either. Studies in Germany (Sprietsma, 2013) and in the Netherlands (van Ewijk, 2011) show that the same essays when signed by a name typical for immigrant societies are likely to get lower grades from primary school teachers. Most, if not all, cognitive/academic achievement tests seem to have some form of bias, either due to the way the test material is prepared/presented, or due to the biased views of individual human evaluators.

Not only humans, but computer systems also exhibit bias (Friedman and Nissenbaum, 1996). A number of recent studies demonstrated that machine learning methods also learn the biases in their training data, and proposed ways of mitigating the bias (e.g., Caliskan et al., 2017; Kiritchenko and Mohammad, 2018; Sun et al., 2019; Bender and Friedman, 2018). Clearly, both data source and the methods are susceptible to bias.

Conceivably, however, the automatic systems may exhibit less bias than humans in this task, since the systems seem to learn (even amplify) the biases when there is a strong bias (Zhao et al., 2017). On the other hand, Sprietsma (2013) report that the biases observed in their study is 'weak', in the sense that the bias observed stems from scores assigned by only a minority of the teachers. The scores from the majority of the teachers in their study does not show any clear bias. A system learning from such a low-bias data set may in fact result in automatic systems that are on average less (severely) biased than the human evaluators. The question, however, is an empirical question. Even though the present systems are not yet mature enough to be applicable in the real world, preventing bias in these systems can only be achieved by studying them carefully and responsibly. In particular, focusing on careful analyses of the systems rather than the usual strong focus on the state-of-the-art performance scores in the field is important to understand consequences of using similar systems in real-world applications.

[4]There has at least been a heated debate on the corpora list (https://mailman.uib.no/public/corpora/2019-December/thread.html) upon announcement of the shared task.

6 Summary and Outlook

We described simple (linear) systems for the GermEval 2020 task 1, 'classification and regression of cognitive and motivational style from text'. The systems achieved strong results in the competition. Besides describing the systems, we present a few additional experiments which may help better understanding of the task at hand and the particular solutions used in this study.

Despite the clear signal in the data for predicting the scores relevant to academic success, the results are unlikely to be applicable in practice. Better modeling practices discussed above are likely to improve the success of the systems. However, it is equally important to analyze the methods and understand their strengths and weaknesses. In particular, for such an application with potentially high impact on the society, the biases that may come from the labeled or unlabeled data sets should be identified, and mitigated.

The systems described here probably owe part of their success to the specific texts obtained on a relevant test. An interesting question to investigate is if the same or similar results can be obtained other, more general, types of texts, for example essays written in the school or linguistic output of the authors during more informal communication.

Like many shared task participations, the present study focused mainly on improving the scores. However, another interesting direction for future work is to analyze the models carefully to get further insights into (linguistic) features in the data that correlate with academic success.

References

Federica Barbieri. 2008. Patterns of age-based linguistic variation in American English. *Journal of sociolinguistics*, 12(1):58–88.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Fabio Celli, Fabio Pianesi, David Stillwell, Michal Kosinski, et al. 2013. Workshop on computational personality recognition (shared task). In *Proceedings of the Workshop on Computational Personality Recognition*.

Çağrı Çöltekin, Taraka Rama, and Verena Blaschke. 2018. Tübingen-Oslo team at the VarDial 2018 evaluation campaign: An analysis of n-gram features in language variety identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 55–65.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

Semire Dikli. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, 5:1:4–36.

Reyn van Ewijk. 2011. Same work, lower grade? student ethnicity and teachers' subjective assessments. *Economics of Education Review*, 30(5):1045–1058.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347.

Tommi Jauhiainen, Krister Lindén, and Heidi Jauhiainen. 2019. Language model adaptation for language and dialect identification of text. *Natural Language Engineering*, 25(5):561–583.

Arthur R Jensen. 1980. *Bias in mental testing*. Free Press, New York.

Dirk Johannßen and Chris Biemann. 2018. Between the lines: Machine learning for prediction of psychological traits – a survey. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 192–211. Springer.

Dirk Johannßen, Chris Biemann, and David Scheffer. 2019. Reviving a psychometric measure: Classification and prediction of the operant motive test. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 121–125, Minneapolis, Minnesota. Association for Computational Linguistics.

Dirk Johannßen, Chris Biemann, Steffen Remus, Timo Baumann, and David Scheffer. 2020. GermEval 2020 task 1 on the classification and regression of cognitive and emotional style from text: Companion paper. In *Proceedings of the 16th Conference on Natural Language Processing (KONVENS 2020)*.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, pages 6300–6308.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Julius Kuhl. 2013. *Auswertungsmanual für den Operanten Multi-Motiv-Test OMT*. IMPART-Test-Manuale. Sonderpunkt-Verlag.

Julius Kuhl and David Scheffer. 1999. Der operante Multi-Motiv-Test (OMT): Manual. Technical report, Universität Osnabrück.

Kim Luyckx and Walter Daelemans. 2008. Personae: a corpus for author and personality prediction from text. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Dong Nguyen, Dolf Trieschnigg, A Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd*

international workshop on Search and mining user-generated contents, pages 37–44. ACM.

James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on Reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12, Brussels, Belgium.

Barbara Plank and Dirk Hovy. 2015. Personality traits on Twitter –or– how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. In *International Conference on Weblogs and Social Media*, pages 102–108.

J Philippe Rushton and Arthur R Jensen. 2005. Thirty years of research on race differences in cognitive ability. *Psychology, public policy, and law*, 11(2):235.

David Scheffer and Julius Kuhl. 2006. *Erfolgreich motivieren: Mitarbeiterpersönlichkeit und Motivationstechniken*. Hogrefe Verlag.

Maresa Sprietsma. 2013. Discrimination in grading: Experimental evidence from primary school teachers. *Empirical economics*, 45(1):523–538.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Nianheng Wu, Eric DeMattos, Kwok So, Pin-zhen Chen, and Çağrı Çöltekin. 2019. Language discrimination and transfer learning for similar languages: Experiments with feature combinations and adaptation. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63, TOBEFILLED-Ann Arbor, Michigan.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Çağrı Çöltekin and Taraka Rama. 2018. Tübingen-Oslo at SemEval-2018 task 2: SVMs perform better than RNNs at emoji prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 34—38, New Orleans, LA, United States.

A Appendix

Hyperparameter	range
SVM margin parameter 'C'	$(0.0, 2.0]$
L2 normalization strength	$(0.0, 50.0]$
Maximum char n-grams	$[2, 9]$
Maximum word n-grams	$[2, 5]$
Case normalization	words or none
Scale of transfer features	$(0, 2.0)$

Table 2: The ranges of hyperparameters used during tuning. See Section 3 for detailed descriptions of the parameters and the tuning process.