Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

# The HUIU Contribution to the GermEval 2019 Shared Task 1

**Melanie Andresen, Melitta Gillmann, Jowita Grala, Sarah Jablotschkin,**
**Lea Röseler, Eleonore Schmitt[†], Lena Schnee, Katharina Straka,**
**Michael Vauth, Sandra Kübler[‡], Heike Zinsmeister**
Universität Hamburg, [†]Universität Bamberg,[‡]Indiana University
{melanie.andresen,melitta.gillmann,sarah.jablotschkin}@uni-hamburg.de
{Jowita.Grala,Lea.Roeseler}@studium.uni-hamburg.de
{Lena.Schnee,Katharina.Straka}@studium.uni-hamburg.de
{michael.vauth,heike.zinsmeister}@uni-hamburg.de
eleonore.schmitt@uni-bamberg.de, skuebler@indiana.edu

## Abstract

In this paper, we present the HUIU system for the GermEval 2019 shared task 1. Our system uses linear SVMs with word and POS unigrams and the number of authors as features. We obtain a micro-averaged F-score of 80.67 on the test data, thus ranking 15th out of 19 submissions, or 9th out of nine groups.

## 1 Introduction

This paper describes the contribution of the HUIU team to the shared task on hierarchical classification of book blurbs at GermEval 2019 (Remus et al., 2019). The task is a multi-label classification that assigns categories to books. The labels constitute a hierarchy, i.e., there are several sub-labels to each category. Two tasks were offered: one focusing on assigning more general labels, the second one focusing on additionally assigning finer grained, hierarchical labels. Our team participated in the first task on assigning general labels, i.e., our system assigns each book one or more labels from the following set: 'Architektur & Garten' (Architecture & Gardening), 'Ganzheitliches Bewusstsein' (Holistic Awareness), 'Glaube & Ethik' (Belief & Ethics), 'Kinderbuch & Jugendbuch' (Books for Children and Young Adult Readers), 'Künste, Literatur & Unterhaltung' (Arts, Literature & Entertainment), 'Ratgeber' (Counseling), and 'Sachbuch' (Nonfiction).

The HUIU system was developed as a class project at the University of Hamburg, i.e., all authors participated in a 6-day compact course that provided an introduction to machine learning for linguists and digital humanities researchers, under the supervision of Kuebler and Zinsmeister. All participants had some experience in programming, but only one of the participants had had prior experience with machine learning. This project was intended to provide a practical introduction to machine learning and to familiarize the participants with every step in the process of translating a problem into a machine learning problem, deciding on a machine learning algorithm, a feature set, extracting features, running machine learning experiments, and evaluating the outcomes. The team submitted a contribution to this shared task as well as to the GermEval 2019 shared task 2 (Andresen et al., 2019).

Because of the setting in a short compact course, the team decided to focus on standard machine learning algorithms available in scikit-learn (Pedregosa et al., 2011), with a fairly basic feature set and initially reducing the problem to a single label classification system. We then extended the feature set only minimally, and used a simple method to extend the classification approach towards a system where we can assign at most three labels. Also because of the course setting, we decided that we would not experiment with deep learning architectures.

The remainder of the paper is structured as follows: Section 2 discusses related work, section 3 describes our experimental setup, including the data set, the machine learning experiments, and the evaluation metrics. Section 4 shows the official results, and we discuss additional results on the development set: experiments to determine good settings for our thresholding approach to multi-label classification and a feature ablation study. We conclude in section 5 and discuss future work.

## 2 Related Work

Multi-label classification has not received much attention in the field of Computational Linguistics. The few exceptions concern work in the fields of offense detection (e.g. Ibrohim and Budi, 2019), relation detection (e.g. Surdeanu et al., 2012), and the prediction of medical codes in clinical notes (Mullenbach et al., 2018). These tasks are similar to our problem in that the number of labels differs per instance. For example, each tweet may contain abusive and/or hate speech and the latter may be related to one or to several issues such as creed, sexual orientation, or disability. Similarly, each sentence may contain a wide range of different relations, and each clinical note may contain a different number of medical codes. An interesting case is presented by Chalkidis et al. (2019), who have annotated legal texts with about 7 000 concepts from the European Vocabulary (EUROVOC). This does not only present a case of an extreme multi-label classification, but it also requires few-shot or one-shot learning approaches since most of these concepts are used very infrequently in the texts.

El Kafrawy et al. (2015) present an overview of methods for addressing multi-label classification and ranking. For multi-label classification, they distinguish between methods that transform the problem into single-label classification, adaptations of single-label classifiers, and ensemble methods. Problem transformations consist of sets of 1-vs-all classifiers, 1-vs-1 classifiers, or creating all combinations of labels and treating them as single labels. For classifier adaptation, neural networks are ideal since every label can be represented as a single output node, and depending on their activation level, multiple levels can be chosen, but other methods can be adapted as well. El Kafrawy et al. (2015) come to the conclusion that ensembles of classifiers work best in a multi-label classification situation.

# 3 Experimental Setup

## 3.1 Data Set

We use the data set provided by the shared task. It consists of a training set (containing 14 548 book blurbs), a development set (containing 2 079 book blurbs), and a test set (containing 4 157 book blurbs). For the final submission, we trained on the combination of the training plus the development set. For the additional experiments described in section 4.2, we trained on the training set and evaluated on the development set. Figure 1 shows an example of a book entry, reduced to the relevant fields.

Since we started with a single label classification system, we first used only the first label assigned to a book in the training data. However, the training data may contain more than one label per book. Therefore, we decided to add one training instance per label. I.e., a book with three labels would contribute three training instances, each being assigned one of the labels.

```
<book date="2019-01-04" xml:lang="de">
<title>Die Essenz der Lehre Buddhas</title>
<body>Klar und verständlich führt der Dalai
Lama in die buddhistische Lehre ein und
eröffnet praktische Wege für alle, die
Gelassenheit und inneren Frieden suchen.
Wer diese einfachen, aber bewährten
Grundsätze des Dalai Lama übernimmt
und nach ihnen lebt, der lebt auch in
Harmonie mit sich und seinen Mitmenschen
dies ist die Essenz der Lehre Buddhas.
</body>
<categories>
<category>
<topic d="0">Glaube & Ethik</topic>
</category>
<category>
<topic d="0">Ganzheitliches Bewusstsein
</topic>
</category>
</categories>
<authors>Dalai Lama</authors>
<isbn>9783453702479</isbn>
</book>
```

Figure 1: Example of a book blurb.

## 3.2 Extracted Features

We extracted word and part of speech (POS) $n$-grams as well as the number of authors as features. For the $n$-grams, we used the title and the body of the text, as delineated in the XML (see Figure 1 for an example). We then performed minimal tokenization via a script. For POS tagging, we used *TnT* (Brants, 1998), trained on the Tübingen Treebank of Written Language (TüBa-D/Z) (Telljohann et al., 2006), version 10.

For words and POS tags, we experimented with $n$-grams of length 1-3. In the final system, only unigrams were used as features since bigrams and trigrams negatively affected the results of the classifier. In addition to word and POS unigrams, we used the number of authors as a feature, using the number of commas as indicator of the number of authors.

## 3.3 Methodology

We used scikit-learn (Pedregosa et al., 2011) for our experiments. An initial investigation comparing SVMs (Support Vector Machines) and Random Forest classifiers showed that a linear

| Rank | Team | Subset Acc. | Recall | Precision | micro-F |
|---:|---|---:|---:|---:|---:|
| 1 | Ericsson Research | 83.64 | 89.23 | 84.32 | 86.70 |
| 15 | HUIU | 75.63 | 80.63 | 80.72 | 80.67 |

Table 1: The official results of the HUIU system in comparison to the best performing system.

SVM gave the best results on the development set (in the single-label setting). For this reason, we only report experiments with the linear SVM. A non-exhaustive parameter search reached the best results using the default settings (penalty=l2, loss=squared_hinge, dual=True, tol=0.0001, C=1.0, multi_class=ovr, fit_intercept=True, intercept_scaling=1, class_weight=None, verbose=0, random_state=None, max_iter=1000).

### 3.4 From Single-Label to Multi-Class Classification

Since SVMs are inherently binary classifiers, they internally already split the problem into multiple classification steps. The linear SVM implementation in scikit-learn follows liblinear and implements a 1-vs-all strategy. We decided to use the internal results of the SVM by looking at the decision function provided for linear SVMs to decide whether we should add a second or third label. We used a manually determined threshold of the difference between the probability of the first and second label (or between the second and third respectively). Our best results are based on allowing a second label only and setting the threshold to $\leq 0.19$. For a closer look at the effects of setting thresholds and using multiple labels, see section 4.2.

### 3.5 Evaluation

For evaluation, we used the official scorer provided by the shared task. It reports precision, recall and the micro-averaged F-score, along with subset accuracy (i.e., the percentage of instances that were assigned the correct set of labels). The micro-averaged F-score serves as the main ranking function in the shared task.

## 4 Results

### 4.1 Official Shared Task Results

9 teams had submitted an overall number of 19 results. The HUIU contribution was ranked no. 15, or 9th group. Table 1 shows the HUIU official results in comparison to the best system. Our system is based on word and POS unigrams and the number of authors as features, allowing up to two labels.

The results show that our results reach a micro-averaged F-score that is about 6 percent points lower than the best ranked system.

### 4.2 Additional Results

In this section, we report on additional experiments, where we evaluated on the development set. In an investigation the required number of labels, we use word and POS $n$-grams, but only create one instance per book in the training data, using the first label. In the ablation study, we start with the full system and then systematically take away options.

However, note that the ablation results need to be taken with a large grain of salt since different runs of the SVM with the same setting often result in larger differences than the differences between settings[1]. The settings where we use one label per book seem to be stable, thus the experiments for determining the best number of labels are run only once per setting. The ablation experiments were run twice, and we report the averages. Ideally, every setting should be run several times, but the time constraints of this project did not allow such a procedure.

#### 4.2.1 Number of Labels and Thresholds

Table 2 shows the results when we vary the number of permissible labels from 1 to 3, and it shows the effects of choosing corresponding thresholds. The threshold is defined as the difference between the probability the SVM assigns to the first and the second label (or the second and third respectively) in the internal 1-vs-all binary classifications. I.e., if we have a high threshold, corresponding to a large difference between the probabilities of the two labels, the system is very permissive in choosing a second label. If the threshold/difference is low, the first and second label need to be very close in probability for the second label to be added.

The results show that there are small differences in the F-score when allowing different numbers of

[1] The cause for this large variation is unclear.

| Setting | Threshold | Subset Acc. | Recall | Precision | micro-F |
|---|---|---|---|---|---|
| 1 label | n/a | 76.48 | 76.95 | 82.54 | 79.65 |
| 2 labels | 1.0 | 68.40 | **83.50** | 74.84 | 78.93 |
| | 0.3 | 72.10 | 81.52 | 77.86 | 79.65 |
| | 0.2 | 73.64 | 80.40 | 79.20 | 79.80 |
| | 0.19 | 73.88 | 80.31 | 79.49 | **79.90** |
| | 0.15 | 73.93 | 80.09 | 79.48 | 79.79 |
| | 0.1 | 74.80 | 78.74 | 80.55 | 79.64 |
| | 0.05 | **75.52** | 77.80 | **81.34** | 79.53 |
| 3 labels | 0.19; 0.2 | 73.79 | 80.81 | 78.62 | 79.70 |
| | 0.19; 0.19 | 73.79 | *80.85* | 78.70 | 79.76 |
| | 0.19; 0.18 | 73.79 | *80.85* | 78.73 | 79.78 |
| | 0.19; 0.17 | 73.79 | 80.76 | 78.78 | 79.76 |
| | 0.19; 0.15 | 73.79 | 80.76 | 78.85 | *79.80* |
| | 0.19; 0.12 | 73.88 | 80.63 | *78.96* | 79.79 |

Table 2: Results when varying the number of permissible labels and thresholds (on the development set).

| Setting | Subset Acc. | Recall | Precision | micro-F |
|---|---|---|---|---|
| full version | 74.73 | 81.17 | 80.20 | 80.68 |
| no author | 73.65 | 79.94 | 78.65 | 79.29 |
| no POS | 75.20 | 81.57 | 80.74 | **81.16** |
| no author/POS | 74.97 | **81.59** | 80.55 | 81.07 |
| no author/POS/title | 74.22 | 80.70 | 80.01 | 80.40 |
| no author/POS/title; one instance | 73.88 | 79.96 | 79.70 | 79.83 |
| no author/POS/title; one instance/label | **76.86** | 77.31 | **82.92** | 80.02 |

Table 3: Results of the ablation study (on the development set).

labels: When we use only one label, we reach an F-score of 79.65, the best result using two labels reaches 79.90, thus giving us a minor boost in performance. Surprisingly, subset accuracy is also highest when allowing only one label. Allowing a third label results in an optimal F-Score (for this setting) of 79.80, i.e., it does not reach the highest F-score when using two labels.

However, when we look at the precision and recall scores, we see a different picture: Using one label gives a high precision but rather low recall, which is understandable since all books that have more than one label in the gold standard will at best be classified only partially. However, this setting also reaches the highest subset accuracy. Adding a second label with a high threshold of 1.0 reverses this picture, i.e., we gain in recall by adding more labels, but precision suffers. The more we lower the threshold the more we lose in recall but gain in precision. Thus, we need to find a good balance for the threshold.

#### 4.2.2 Ablation Study

Table 3 shows the results of our ablation experiments. We start with the full system that also served as the basis for the official submission. We see that leaving out the number of authors results in a minor deterioration, but leaving out the POS information results in a boost in F of about 0.4, equally distributed across precision and recall. We had originally decided to use POS unigrams since they improved results in the single-label setting. This shows that the ideal settings do not transport across single-label and multi-label experiments.

Leaving out both author and POS information results in a minimal loss, leaving out the title information and using only one instance per book with the first label result in a smaller loss. Restricting the system to a single-label task results in a minimal improvement in F, based on high precision, but low recall. Surprisingly, this setting also provides the highest score for subset accuracy.

## 5 Conclusion and Future Work

This project was mostly carried out in the setting of a 6-day compact course. Given the time constraint, we have shown that we can put together a fairly robust system for multi-class classification of books into categories. Our system ranked about 6 points below the best performing system.

Future work should investigate using additional features, such as looking into sentence length, the syntactic complexity of sentences, or the occurrence of named entities. We also need to investigate the issue of variation in the SVM results when we use more than one instance per book while there is no variation at all when we only use one instance. Another point is to investigate ensembles as suggested by El Kafrawy et al. (2015).

## References

Melanie Andresen, Melitta Gillmann, Jowita Grala, Sarah Jablotschkin, Lea Röseler, Eleonore Schmitt, Lena Schnee, Katharina Straka, Michael Vauth, Sandra Kübler, and Heike Zinsmeister. 2019. The HUIU contribution to the GermEval 2019 shared task 2. In *Proceedings of the GermEval 2019 Workshop*, Erlangen, Germany.

Thorsten Brants. 1998. *TnT–A Statistical Part-of-Speech Tagger*. Universität des Saarlandes, Computational Linguistics, Saarbrücken, Germany.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. Extreme multi-label legal text classification: A case study in EU legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, MN.

Passent El Kafrawy, Amr Mausad, and Heba Esmail. 2015. Experimental comparison of methods for multi-label classification in different application domains. *International Journal of Computer Applications*, 114(19).

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the ACL*, pages 1101–1111, New Orleans, LA.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Steffen Remus, Rami Aly, and Chris Biemann. 2019. GermEval-2019 task 1: Shared task on hierarchical classification of blurbs. In *Proceedings of the GermEval 2019 Workshop*, Erlangen, Germany.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, Jeju Island, Korea.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, and Heike Zinsmeister. 2006. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Germany.