Deriving Players & Themes in the Regesta Imperii using SVMs and Neural Networks

Juri Opitz and **Anette Frank**
Department of Computational Linguistics
69120 Heidelberg, Germany
{opitz|frank}@cl.uni-heidelberg.de

Abstract

The Regesta Imperii (RI) are an important source for research in European-medieval history. Sources spread over many centuries of medieval history – mainly charters of German-Roman Emperors – are summarized as "Regests" and pooled in the RI. Interesting medieval demographic groups and players are i.a. *cities, citizens* or *spiritual institutions* (e.g. bishops or monasteries). Themes of historical interest are i.a. *peace and war* or the endowment of *new privileges*. We investigate the RI for important *players and themes*, applying state-of-the-art text classification methods from computational linguistics. We examine the performance of different classification methods in view of the linguistically very heterogeneous RI, including a Neural Network approach that is designed to capture complex interactions between players and themes.

1 Introduction

The Regesta Imperii (RI)[1] are considered a fundamental, autonomous source for German and European history. It extends over many centuries, from the Karolinger dynasty to Maximilian I, from around 800 to 1500 AD. The RI have their roots in the 19th century, when the German librarian Johann Friedrich Böhmer started to collect and document the charters (including known and possibly unknown fakes) of the German-Roman emperors, in terms of so-called *Regests*. The Regests contain relevant judicial content of the referenced charters (cf. Zimmermann (2000), Niederkorn (2005), Rübsamen and Kuczera (2006)). A royal charter was created, for example, when an emperor decided to give a land grant, or privileges such as new rights to one of his landlords or cities.

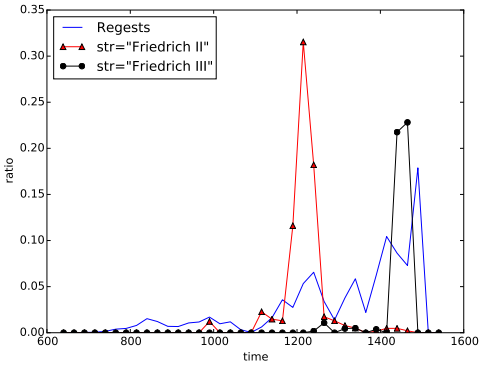

Figure 1: Blue line: distribution of the 129,504 Regests across time. Others: ratios of Regests, in which the terms "Friedrich II." (triangles) and "Friedrich III." (circle) occur. The names of these German-Roman kings are examples for concepts which are rather confined in time in the RI.

Covering about 13 million tokens, the RI constitutes a large-scale resource that is still growing today[2]. The 129,504 Regests we have access to can be treated as a collection of corpora (e.g., one corpus for each Roman-German emperor dynasty), or as a single corpus covering all collected materials. Our work treats the RI as a single corpus. The RI comprises texts written in different German varieties, as well as Latin. Often we find up to three different languages or varieties within a single Regest.

As seen in Figure 1, the Regests are not evenly distributed over time but have the greatest mass from about 1200 to 1500 AD. Many terms and concepts only occur in certain times. An overview

[1] http://www.regesta-imperii.de/cei

[2] We retrieved a "snapshot" of the RI via the public REST interface http://www.regesta-imperii.de/cei/ on 26.4.2015.

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 74–83,
Berlin, Germany, August 11, 2016.

# Regests	129,504
# types	$\approx$ 407,000
# tokens	$\approx$ 13,000,000
mean length (in tokens)	$\approx$ 85
median length (in tokens)	$\approx$ 52
ttr_{log}	$\approx$ 0.79
ttr_{log} SDeWaC	$\approx$ 0.68

Table 1: Corpus statistics for the RI at the time we used it. $\text{ttr}_{log} = \frac{log(\#types)}{log(\#tokens)}$ is the logarithmic type-token ratio. Taking the logarithm allows better comparison with corpora of different sizes. SDeWaC is a German Corpus comprising 44 million sentences crawled from the internet.

abbreviation	groups and themes traced in RI
b0	nobility, nobles
b1	spiritual Institutions
b2	lesser nobles
b3	city, citizens
b4	Jews
b5	women
b6	new privileges
b7	confirmation of privileges
b8	land grants, land bestowal
b9	finances
b10	justice
b11	war and peace

Table 2: Traced demographic groups and themes.

of corpus statistics is given in Table 1. The high logarithmic type-token ratio (ttr_{log}) supports the observation that the language of the RI is highly heterogeneous: although the domain of the RI is rather focused (abstracts of medieval charters), it is notably higher than what we find in the contemporary German SDeWaC corpus[3].

A Regest itself is a very unique form of a document, and some of them are not easy to comprehend even for humans. Consider

Example 1 *A Regest from 1332, issued in Parma by Karl IV. (*1316, †1378).*[4]

> bekennt dem Johann de Landulphis, iudici et auditori curie paterne et sue, achtzig goldgulden für besoldung und sechzig goldgulden wegen versendungen desselben schuldig zu sein. Registr. priv. von Pavia hs. (fol. pap. sec. 15 vel 16) zu Pavia bl. 5.

The Regest describes an action of King Karl IV. in 1332, in Parma, Italy. Karl IV. acknowledges that he owes "Johann de Landulphis", "achtzig goldgulden" (eighty gold coins) for wages and "sechzig goldgulden" (sixty gold coins) for reasons which are rather difficult to interpret: "(...) wegen versendungen desselben schuldig zu sein" (interpretable as wages and travel expenses). Beyond that, the Regest contains information in Latin ("iudici et auditori curie paterne et sue"), plus references and meta information (last sentence).

It is easy for humans to infer that the theme of the above Regest is about *finances* (indicated by mentions of "goldgulden" (gold coins) and "besoldung" (wages)). Further, a specific group of persons plays a role, namely *nobles*. This is indicated by "de" in the name of "Johann de Landulphi", who is promised money by the king. The Latin "de" in the middle of a name generally suggests that the person belongs to the class of nobles, as in "Elizabeth of (=de) England". So, one may conclude that in the above Regest, the players are *nobles*, acting under the theme *finances*.

Our aim is to trace within the RI interesting *demographic groups* joint with the *themes of their interactions*. We aim to identify which Regest is about which theme(s) and group(s), to perform interesting data analysis, e.g. visualizing the importance of different groups and themes not only in relation to time but also in relation to other factors such as issuer, location, and possibly more.

With the support of a domain expert we determined interesting demographic groups (players) and themes which play a role in the Regests. All players and themes can be treated as individual binary classification problems. An overview is given in Table 2. It can be interesting, e.g., to relate the occurrence of *city* or *citizens* with occurrences of *privileges* with respect to time, thus approximately tracing the development of *privileges* for cities.[5]

A Regest can be labeled with zero to all of the 12 selected labels. Thus, there exist many possible combinations.[6] We cast the labeling problem as a multi-label document classification task, allowing several labels (i.e. groups and themes) to be assigned for a single document (i.e. Regest).

For automatic pattern recognition on this historic data, we deploy four state-of-the text classification methods, (i.) Support Vector Machines (SVM) (binary classification); (ii.) Semi-

[3] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/sdewac.html

[4] cf.: RI VIII n. 1, in: Regesta Imperii Online, URI: http://www.regesta-imperii.de/id/1332-09-22_1_0_8_0_0_7_1(29.04.2016).

[5] This is an important field of historic research because rights of European cities originated in the Middle Ages.

[6] Since each group and theme represents a binary variable, there are 2^{12} possible combinations.

Supervised SVMs (S^3VMs), to exploit the large amount of unlabeled data; (iii.) a Neural Network as a meta-learner applied to the SVM outputs (do the groups and themes influence each other?) and (iv.). a Convolutional Neural Network (CNN) classifier with pre-trained word vectors as input, which operates directly on the input documents.

We evaluate all methods on a manually labeled test set and perform data analyses on the full RI to illustrate its usage in Digital Humanities research.

2 Related Work

To the best of our knowledge, no (published) research has yet been conducted in the Digital Humanities community about NLP of the RI. Kuczera (2015) experimentally transfers attributes and relations between entities from the times of Friedrich III. (i.e. a subset of the RI) into a graph database and shows how historians could profit from the possibilities offered by such structured data repositories.

Ruotsalo et al. (2009) suggests that knowledge- and machine learning based NLP methods can help with complex annotation tasks in the cultural heritage domain. Their experiments demonstrate that automatic annotation of certain roles in artwork descriptions closely matches the performance of human annotators.

Piotrowski (2012) gives an overview of the manifold challenges in applying NLP to historical documents. He reports that the effectiveness of normalization strongly depends on text type and language, and satisfying results are achieved mainly on more recent texts. Piotrowski concludes that "the highly variable spelling found in many historical texts has remained one of the most troublesome issues for NLP". Thus we chose our procedure to not depend on normalized texts.

Massad et al. (2013) give an overview of the processing of recorded history texts. They examined a graph-based approach and an approach based on NLP. In their NLP experiments they analyzed the Wikipedia corpus with respect to time, relating specific strings and n-grams to time and page edits. The authors suggest that future research should focus on data analysis, trends and, most importantly, the access to historic corpora spanning a larger time span compared to the corpus employed in their experiments. We think that our research covers these aspects.

Meroño-Peñuela et al. (2014) in their survey on *History and Computing* propose NLP methods for dealing with raw corpora, yet do not propose specific tools due to manifold decisions to be taken, that strongly depend on the nature of the data.

3 Approach

The aims of our work are two-fold. On the application side, we aim to discover structures involving players and themes over times in the RI. On the methods side, we investigate to what extent Neural Networks (NN) are capable of learning complex relationships between players and themes, beyond the capacity of ordinary SVM classifiers that treat each classification label independently. E.g., if *nobles* play a role in a given Regest, it seems more likely that it is about *bestowal of land*, rather than e.g. justice, which presumably concerns other groups equally. We compare two architectures: a NN that builds on the output of *independent* binary SVM classifiers, in addition to other information, such as document vectors, in contrast to a full-fledged Convolutional Neural Network (CNN).

3.1 Preprocessing

Given the heterogeneous nature of the RI, we do not perform major pre-processing of the data. The Regests are only tokenized and converted to lower case. Thereafter they are mapped to boolean and tf-idf vectors of dimensions 2,000 and 10,000. The value at index i of a boolean vector representing document d encodes whether the term represented by i appears in d (1) or not (0). Tf-idf is similar but assumes that words that appear in many documents are less informative, and hence their respective vector-value is decreased[7].

3.2 Using SVMs and S^3VMs

SVMs are binary maximum-margin classifiers that can be extended to the multi-label case by training one SVM for each label. Semi-supervised SVMs (S^3VMs) work by forcing the hyperplane separating the labeled data with margin also through low density regions of space, making use of the cluster hypothesis (Chapelle et al., 2008). S^3VMs have been shown to be very successful especially when few labeled training data is available (Sindhwani and Keerthi, 2006). The downside is that

[7] More precisely, it is weighted by term frequency and the logarithm of $\frac{|D|}{|D_w|}$, where $|D|$ is the number of all documents and $|D_w|$ the number of documents in which w appears.

the S^3VM's optimization problem loses the global optimality of the standard SVM problem.

In both approaches – SVM and S^3VM – the assumption is that labels do not influence each other. I.e., if *women* play a role in a given Regest, it is *not* less likely that it is also about *war and peace*.

3.3 Combining SVMs and NNs

To enable our classifiers to capture possible dependencies between players and themes, we extend the SVM classifiers with a Neural Network, realizing a meta-learning architecture. The NN may learn that if groups x and y participate in a Regest, some theme z is unlikely to occur, even if predicted so by an independent binary classifier.

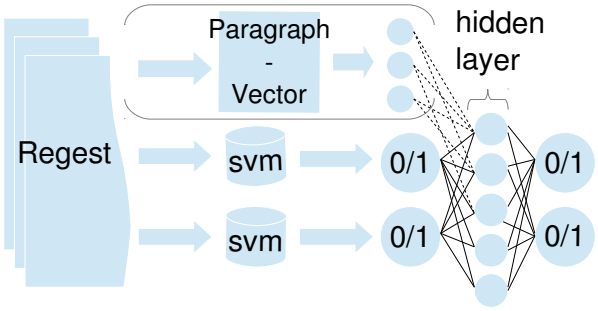

Figure 2: A Neural Network as a meta learner over multiple binary classifier's outputs, supplemented with a paragraph vector over the document. The figure is simplified to a multi-labeling task with only 2 classifiers (in reality there are 12 SVMs).

After choosing the "best" SVMs for each label, the outputs of the SVMs are fed into a Deep Neural Network (cf. Figure 2). We employ three input settings: (a.) using SVM output labels only, (b.) using SVM output labels and the document vectors (the boolean variant), and (c.) the SVM output labels jointly with *Paragraph Vectors*. Paragraph Vectors are learned similar to word embeddings but represent sentences or documents. They have been shown to yield strong performance in classifying sentences, IMDB opinions and also in Information Retrieval. As the Regests are short documents, they are suitable for being represented by these dense vectors, which are learned in an unsupervised manner (Le and Mikolov, 2014).

3.4 Using Convolutional NNs

Recently, CNNs have been successfully applied to various text and semantic sentence classification tasks, and often achieved very good performance (Kim, 2014; Zhang et al., 2015). Since CNNs usually require large numbers (thousands or more) of training samples to achieve very good performance, it would come rather as a surprise if trained on some few hundred samples, they would generalize better on unseen data compared to a mix of binary maximum-margin classifiers. We included this setting to serve as a baseline on the Neural NLP side and generated pre-trained word-embeddings of two sizes using all Regests.

4 Experiments and Results

4.1 Experimental Setup

Training and Test Data. We manually labeled 500 Regests, randomly drawn from the corpus to prevent bias.[8][9] The data was split into a training and test section of 400 and 100 Regests. The first two lines in Table 4 display the distribution of players and themes in the annotated data. Some of them occur rarely in both training and test data (e.g. Jewish people (b4) with only 3% and 2% of the respective data sets). On the other hand, *nobles* play a role in over 70% of annotated Regests. For estimation of model parameters we apply cross validation (CV) on the training set. We proceed as follows: (i) *Parameter tuning of SVMs*. For each different vector size and representation scheme, we tune the inner parameters of an SVM with CV on the training data. (ii) *Testing of SVMs*. We retrain each SVM on the full training data using the chosen hyperparameters, and evaluate the model on the test data set. (iii) *Determining an independent multi-label system*. As input to the NN models as meta learners over SVM outputs, we determine an IMC ("Independent Margin Classifiers"), a set of independent margin classifiers, consisting of the 12 SVMs that achieve maximum training CV score for each problem. (iv) *Training NN models.* For different NN models we again determine hyperparameters with CV on the IMC-outputs for the training section, and retrain the final NN models on the full training data, before (v), *Testing of the NNs* is again done on the final test set.

Evaluation Metrics. Our evaluation needs to take into account that many labels underlie a skewed distribution (cf. Table 4). For example,

[8] One of the authors, with experience in history sciences, annotated the data. In future work we plan to obtain comparable annotations possibly with help of experts in online history forums.

[9] Our data set and further details of experimental settings are available at `https://cl.uni-heidelberg.de/~opitz/ri/`

consider that one label only is positive among 100 test samples. A classifier that labels all instances as negative yields a deceivingly high score of 0.99 accuracy. Hence we employ *Balanced Accuracy*, the mean of Recall (Sensitivity) and inverse Recall (Specificity[10]), defined as $Acc_{bal} = \frac{Sensitivity+Specificity}{2}$.

In the above example, where Accuracy yields a biased score of almost one, balanced Accuracy yields a more realistic value of 0.5. Given the unbalanced distribution of our test data set, we report balanced accuracy for each of the 12 binary problems. We also report their arithmetic mean $\overline{Acc_{bal}}$ to provide a global measure of performance.

Baselines. As Baselines we choose, besides a simple majority voter, a Multinomial Naive Bayes algorithm, which is commonly used in text classification tasks (both in an independent binary manner for each label). Table 3 shows that Naive Bayes improves over the majority baseline for all problems and yields a solid 0.67 $\overline{Acc_{bal}}$, 0.17 pp. above the majority voter.

IMC achieves 0.795 $\overline{Acc_{bal}}$ and significantly outperforms both the majority baseline (+0.3 $\overline{Acc_{bal}}$) and Naive Bayes (+0.13). For each problem the score is better with up to +0.47 Acc_{bal} for recognizing *women* (b5) in a Regest. For *lesser nobles* (b2) and *war and peace* (b11), the independent classifiers combination baseline yields the overall best results (0.62 and 0.79 Acc_{bal}).

4.2 Evaluation Results: In Depth Analysis

SVMs/S^{3}VMs combined into the multi-labeler ("IMC", Table 3) achieve good performance (0.795 $\overline{Acc_{bal}}$). Based on the training CV scores, IMC consists of six supervised SVMs and four S^{3}VMs. S^{3}VMs in the IMC were chosen for problems b0, b1, b8 and b10. With respect to b2 and b11, IMC outperforms all NN approaches (b2: +0.04, b11: +0.01). The Naive Bayes Baseline is outperformed with +0.128 $\overline{Acc_{bal}}$. This strong improvement could be due to the generalization capacity of the maximum margin, which might be especially useful with small training set sizes.

With regard to representation schemes such as boolean or tfidf and 2,000 words or 10,000 words, we observe no clear patterns whether one works generally better than the other on the RI. 5 classifiers of IMC are trained on 10,000 words and 10 classifiers use boolean word-features.

[10] $Specificity = \frac{TN}{TN+FP}$

CNNs fed with 128 dimensional embeddings outperform majority vote (+0.06 $\overline{Acc_{bal}}$) but not Naive Bayes (-0.11), most likely due to the low amount of training data. Another explanation is that the 129,504 Regests were not sufficient to pretrain useful word-vectors (possibly also negatively influenced by the word variety). As the vector size increases (512 dimensions), the performance drops further (+0.01 over the majority voter).

The remaining classifier models are intended to detect dependencies between players and themes and had access to the outputs of IMC. Specifically, the question is whether NNs are suitable for detecting such dependencies. As baselines we considered SVM and Decision Tree models, trained on the outputs of the independent learners (in Table 3: +Decision Tree, +SVMs). Neither copes specifically well with this input information (-0.045 $\overline{Acc_{bal}}$ for +Decision Tree and -0.007 for +SVMs). Even when supplied with more information using various sizes of Paragraph Vectors (omitted in Table 3), both systems do not improve their previous scores.

Neural Networks employed as meta learners, by contrast, are able to improve results for specific problems, especially when supplied with Paragraph Vectors, resulting in the overall best system on test, a NN with 2048 hidden nodes and Paragraph Vectors of dimension 512 (+NN$_{2048}$+PV$_{512}$, Table 3). Still, the overall performance gain is small with only +0.004 $\overline{Acc_{bal}}$. When omitting b3 (*lesser nobles*) from the result calculation (it was the most controversial class in the annotation), the gain over IMC increases to +0.006. Notable individual performance gains are achieved for b0 (*nobles*, +0.02), b6 (*new privileges*, +0.05) and *bestowal of land* (+0.02). We conclude that there are dependencies between *nobles*, *bestowal of land* and *privileges* which cannot be captured by considering these classes independently.

To analyze on which groups and themes the neural network meta-learner offers *significantly* differing predictions ("it disagrees with its input"), we calculate mid-p-values with McNemars test (Fagerland et al., 2013) between different systems outputs (cf. Table 3). Comparing the best three NNs among each other, the 1,200 single predictions each system made do not differ significantly (min. $p = 0.065$), however the opposite is true when comparing the best three NNs to IMC (all p-values < 0.05). This indicates that there is more

classifier types	b0	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11	$\overline{Acc_{bal}}$
majority baseline	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Naive Bayes	0.52	0.78	0.62	0.53	0.72	0.51	0.62	0.68	0.71	0.74	0.67	0.69	0.667
SVMs + PV_{128}	0.5	0.51	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.501
IMC	0.72	**0.82**	**0.62**	**0.76**	**1.0**	**0.98**	0.66	**0.94**	0.8	**0.76**	**0.7**	**0.79**	0.795
+Decision Tree	0.68	**0.82**	0.58	**0.76**	0.75	**0.98**	0.63	0.86	0.71	0.75	**0.7**	0.77	0.75
+SVMs	0.72	**0.82**	0.58	**0.76**	**1.0**	**0.98**	0.67	0.94	0.76	**0.76**	**0.7**	0.77	0.788
+NN_{2048}	0.72	**0.82**	0.58	**0.76**	**1.0**	**0.98**	0.68	**0.94**	**0.84**	**0.76**	**0.7**	0.76	0.796
+NN_{2048}+PV_{512}	**0.74**	**0.82**	0.58	**0.76**	**1.0**	**0.98**	**0.7**	**0.94**	0.82	**0.76**	**0.7**	0.77	**0.797**
+NN_{2048}+BV	0.71	**0.82**	0.56	**0.76**	0.5	0.97	**0.7**	0.91	0.55	**0.76**	**0.7**	0.78	0.727
CNN_{128}	0.56	0.49	0.52	0.5	0.5	0.55	0.53	0.61	0.5	0.6	0.64	0.67	0.557

Table 3: Performance of different systems on the test set, measured with balanced accuracy (Acc_{bal}). Majority vote and Naive Bayes represent first-order baselines, IMC can be viewed as a second-order baseline. Systems marked with + have access to individual classifier outputs (IMC) and optionally paragraph (PV) and bag-of-words (BV) vectors. Best scores for each group are bolded. Underscores mark an improvement ≥ 0.03 (3%) Acc_{bal} for a specific group by NN classifiers over the IMC baseline or vice versa. NN_n: Neural Network with n hidden units, PV_n: Paragraph Vectors of dimension n, BV: boolean bag-of-words, CNN_n: Convolutional Neural Network with pre-trained vectors of dimension n.

binary problem	b0	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11
prevalence RI Train	0.56	0.33	0.1	0.23	0.03	0.1	0.15	0.13	0.07	0.18	0.1	0.19
prevalence RI Test	0.53	0.43	0.12	0.15	0.02	0.05	0.2	0.13	0.09	0.2	0.07	0.16
prevalence RI IMC	0.77	0.39	0.17	0.15	0.01	0.07	0.23	0.13	0.16	0.13	0.09	0.26
prevalence RI NN	0.7	0.39	0.07	0.15	0.01	0.07	0.19	0.13	0.06	0.13	0.09	0.18

Table 4: Prevalence of groups and themes: humanly labelled data vs. full automatically labelled RI.

consensus about the label-predictions between different NN architectures than between the NNs and IMC. On the binary problem level, the p-value for theme 6, new privileges, lies below 0.05 for all NN architectures with more than 512 hidden units. For b8, land grants, all p-values are < 0.05 for architectures with more than 128 hidden units. Observation of the predictions further suggests that the NNs feel the most need to correct the SVMs with b6 and b8 (with these the correction ends up in better predictions) and b0, nobles. However, in predicting nobles the difference is never significant. For example, NN2048+PV512, the best NN on the test set disagrees with the SVM on the nobles-label in 11 of 100 cases. Here the NN is correct only in 6 cases, making the difference non-significant with p=0.77. With b8 on the other hand there are 14 disagreements and 13 accurate corrections, resulting in a p-value of 0.001 (b6: 6 corrections, 6 accurate, p=0.016). Taking predictions over all groups again (1,200 predictions), this NN differs in 46 cases from the IMC choice and is correct 39 times ($p < 0.0005$). Why is the resulting perfomance increase in $\overline{Acc_{bal}}$ only 0.2%? This is due to the fact that the NN is more restrictive in assigning labels than the independent learner model: in all 129,504 Regests, it predicts 50,968 less positive labels than IMC. As positive labels are strongly under-represented in the manually labeled data, the (non-weighted) Acc_{bal} measure is much more influenced by an additional True Positive than a True Negative for a rare group or theme.

Paragraph Vectors (PV) used as input to the NNs apparently contain more information than standard (boolean) bag-of-word (BoW) vectors. When the best NN is fed with BoW vectors instead of PVs it achieves lower performance (-0.07 $\overline{Acc_{bal}}$). To test whether Paragraph Vectors work better simply in general, we trained 12 independent SVM classifiers on PVs only, to predict players and themes. The result, for several dimensions of Paragraph Vectors (between 64 and 2048) fed into an SVM (best result: SVMs+PV_{128} in Table 3), did not exceed the Naive Bayes baseline, indicating strongly that PVs alone are inferior to BoW vectors for standard textual classification of the RI. Our explanation is as follows: While Quoc Le (2014) achieved good results in classifying sentiment of movie reviews with Paragraph Vectors, he hypotheses that movie reviews are tailor-cut for learning the vectors for this problem, because *compositionality plays an important*

role in deciding whether the review is positive or negative. The RI are a more complex source and it is debatable whether compositionality plays a role with regard to co-occurring groups and themes. Also, while movie reviews often contain similar (sentiment) vocabulary, each Regest presents its content in rather unique ways. The NN that learns Paragraph Vectors is thus presented with very diverse information, most likely generating vectors containing every and thus little information. We conclude that using standard BoW vectors as first-order information was the correct choice, while PVs prove more suitable as higher-order information for the NN acting as a meta-classifier (as they add little but additional information).[11]

Players and themes that can be predicted with great success by many systems on the test set are *confirmation of privileges* (b7: 0.94), *Jews* (b4: 1.00) and *women* (b5: 0.98). By contrast, all systems fail to reliably predict class b2 (*lesser nobles*), which yields a maximum of 0.12 points beyond majority and no gains beyond Naive Bayes. One explanation for this low performance is that it was really hard (if not sometimes impossible) to distinguish between non-nobles and nobles in the annotation process. All other groups and themes can be predicted with solid accuracy scores ($\geq$ 0.20 above majority, $\geq$ 0.02 above Naive Bayes, and $\geq$ 0.62 Acc_{bal} per category in general).

The system +NN_{2048}+PV_{512} perfoms best in $\overline{Acc_{bal}}$. We also analyze two additional criteria of performance: (i) the Kullback-Leibler (KL) divergence between distributions of labels in the manually annotated data to the distributions of labels automatically assigned to the full RI and (ii), the KL divergence between the distributions of amounts of labels (0-12 labels can be assigned to a Regest). For (i), the KL divergences are $KL_{train,test} = 0.033$ and $KL_{train,RI_{NN}} = 0.036$, $KL_{train,RI_{IMC}} = 0.058$ indicating only a small divergence between human and automatic labeling by the NN w.r.t. the distributions of the twelve groups and themes (cf. Table 4). In fact, all of the best three NNs appear to have smaller KL-divergencies than IMC. Also (ii), number of group and theme labels that are assigned by human vs. automatic labeling, shows similar tendencies: $KL_{train,test} = 0.02$, $KL_{train,RI_{NN}} = 0.01$, $KL_{train,RI_{IMC}} = 0.07$. On average, two labels were assigned to a Regest by all labeling systems. The human assigned 43% of the Regests two labels, IMC 27% and the NN 34%.

In sum, our results indicate that NNs can learn dependencies of labels from independent classifier predictions. NNs are thus suitable to detect structures in the data that are intuitive for humans.

5 Deriving Structures of European Medieval Times

We labeled all Regests with $+NN_{2048} + PV_{512}$. We eye-balled several annotations and found many of the predicted classes to be correctly inferred[12].

5.1 Feature Analysis

The learned weight vectors of the SVMs offer interpretation of the terms w.r.t. the classified groups and themes. Table 5 displays, for selected classes, the phrases which were assigned the highest weights. Many of these intuitively make sense. Indicator terms for War and Peace are "truppen" (troops), "friedensverhandlungen" (peace talks) and the preposition "gegen" (against), other terms point geographically to the East: türken (turkish) or konstantinopel (today: Istanbul).

From the analysis we conclude that the decision to not normalize the texts was reasonable, given that we find many high-weighted terms that are abbreviations, e.g., "urkk" (charter), "kgin" (queen) or latin expressions: 'ecclesia" (christian community), "abbati" (father), "monasterii" (locative of monastery) are indicators for spiritual institutions.

5.2 Investigating the Regesta Imperii

Using the automatically assigned labels for *players* and *themes* in the full set of 129,504 Regests, we are able to investigate structures that emerge between specific players and themes, with respect to time or certain locations. In Figure 3 we trace the development of the ratio of Regests which were both about *cities* and *privileges* w.r.t. time. Given that in some years no Regests are available, the ratios are "smoothed" by calculating them over bins of 25 years. The occurrence ratio is determined by $Ratio(gt, b) = \frac{|Regests_{b,gt}|}{|Regests_b|}$, where gt is the set of groups and themes we want to "trace" and b is one of the bins of 25 years. $|Regests_{b,gt}|$

[11] Note that this applies only to NNs as meta-learners: the SVM-based meta-learner baseline performed below majority baseline when supplied also with Paragraph Vectors ($\overline{acc_{bal}}$ with additional Paragraph Vectors: 0.786, without: 0.788).

[12] The automatic annotations (for the full RI) can be obtained from `https://cl.uni-heidelberg.de/~opitz/ri/`

Group/Theme	highest (positively) weighted terms
Spiritual Institutions	ecclesia, reims, lucius, erzbischof, abbati, imperatoris, abt, nonnenkloster, mönch, bischfliche vice, konvent, intervention, monasterii, episcopi, kloster, besitz, besitzungen, bischof, papst, kirche
Jews	haupt, angesichts, anspruch, aufnehmen, freyburg, niemals, christen, vidimus, heilbronn ungelt frevel, judenschaft, stifter, quittieren, kost, verstoßen, christliche, gebrechen, einnehmer, judensteuer
City/Citizens	docum, beglaubigung, landfriede, weltlich, reichssteuern, cons, gemeinde, breslau, schffen, urkk gelnhausen, laden, verhören, einwohner, rathe, städte, bürgern, brgermeister, bürger, stad, stadt
War and Peace	friedensverhandlung, entschdigung, kräften, schiedsspruch, hoffe, umso, castilien, klar, sehr, türken pabstes, belagerung, dienen, konstantinopel, sagt, friede, truppen, kriege, krieg, gegen
Justice	verhngt, einwohnern, schiedsrichter, aberacht, lichtenberg, gewhrte, theile, bestraft, begangen, stand fremdes, landgerichte, verlorene, landgericht, einerseits, andererseits, wiedergutmachung, urteil

Table 5: Highly weighted terms for groups and themes found in SVM classifiers. Some terms are difficult to translate, but most terms intuitively make sense. For example: many terms for *Jews* relate to financial taxes ("quittieren"–to receipt; "einnehmer"–collector). Other terms for this group are negatively connotated: "frevel"–sacrilege, "verstoßen"–outcast. Jews in medieval times often were at most tolerated and had to pay special taxes (above: "judensteuer"). For all themes and groups a large amount of the heigh-weighted terms is in Latin, suggesting that it was a correct decision not to filter out Latin words.

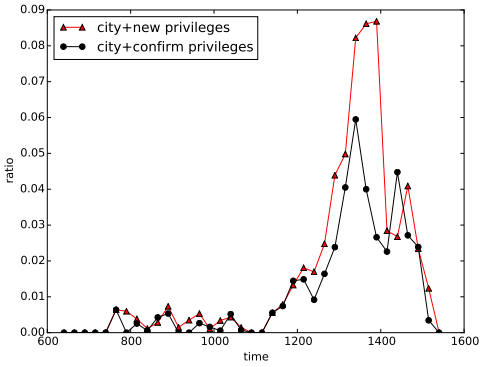

Figure 3: Functions from times to ratios of occurrence for *city* and *new privileges* (triangles) and *city* and *confirmation of privileges* (circles) in the RI. High concentration of *cities* and *privileges* are found from the 12th century onwards, with a peak in the 14th century. *new privileges* outweigh *confirmation of privileges* around the 14th century.

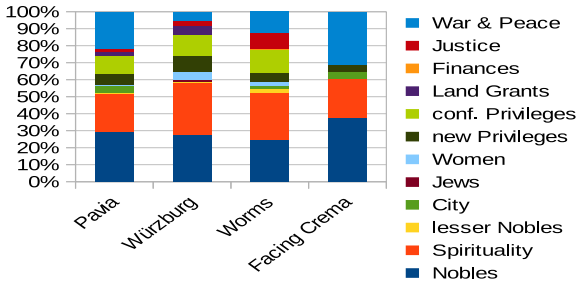

Figure 6: Players and themes in Regests submitted under the name of the German-Roman emperor Friedrich I. (*1122, †1190) in different locations. *War and Peace* played the greatest role in the Italian city Crema, which indeed was involved in war during Friedrich's regency and subjected 1160.

is the number of Regests from time bin b which are about *all* groups and themes contained in gt.

Not only can groups and themes be traced with regard to time, but also to locations or/and to certain emperors. This is exemplified in Fig. 6 and 4 where we count the occurrences of all 12 themes and groups with respect to these parameters and normalize by the sum of all 12 occurrence counts.

6 Conclusions

We solved a multi-label text classification problem to derive interesting demographic groups (e.g. *citizens*) and themes of interactions (i.a. *bestowal of privileges* or *justice*) in the Regesta Imperii.

Evaluation on a held out test set suggests that most groups and themes can be predicted with good reliability: 9 out of 12 classes can be predicted with a (balanced) Accuracy score ≥ 0.75. The arithmetic mean of all 12 scores – our global performance measure – is 0.797 for the system that was finally chosen to label the entire RI.

A Neural Network acting as a meta learner over the outputs of independent maximum margin classifiers and Paragraph Vectors (document embeddings learned by neural networks) led to a minor improvement of 0.2% mean score. However, for the group *nobles* and the themes *bestowal of land* and *new privileges* the scores were improved by up to 3%, 4% and 5%, indicating dependencies between these classes that cannot be captured by classifiers working under the label-independence assumption. We conclude that NNs can give ad-

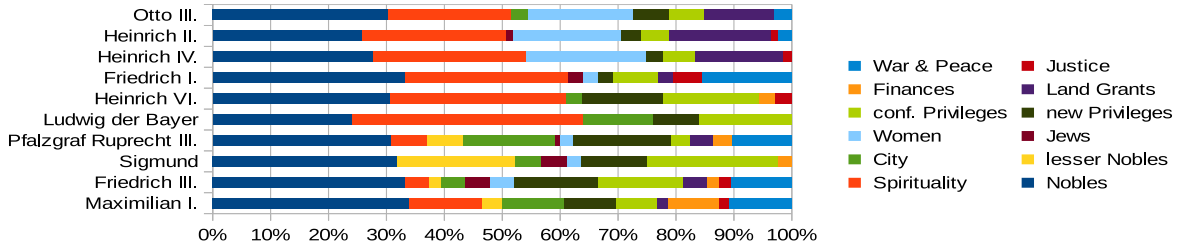

Figure 4: Impact of groups and themes in the German city of Mainz with respect to emperors and time: From Otto III. (crowned German-Roman King in 983) to Maximilan I. (crowned in 1486). The impact of *spiritual institutions* (from Ruprecht III onwards) and *women* and *land bestowals* (both from Friedrich I. onwards) seems to decrease. Finances seem to play a more important role in the later Middle Ages.

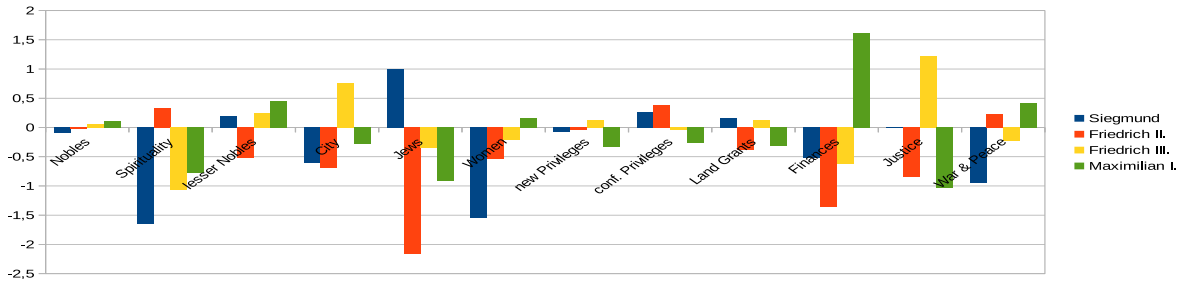

Figure 5: Logistic Regression weights when we force themes and groups to predict the issuing emperor. Negative Weights suggest negative correlation, positive weights suggest positive correlation. Observably *finances* and *war and peace* are associated with Maximilian I. He was notoriously famous for his flamboyant lifestyle and led many wars. Two components leading to great debts, which he mostly owed to Jakob Fugger, banker from the famous Fugger family.

ditional information on possible dependencies between classes in a multi-label classification task.

Conceptually the approach is straightforward, but a complicating factor is the exploding parameter space: Besides the "inner parameters" of the Learners, regularization control or the number of neurons in the Neural Network, there are numerous "outer parameters", e.g., possible ways of document representation or pre-processing.

As best-performing system we determined a NN model with additional Paragraph Vector information. It obtained the best results on the test set and also yields the minimum KL divergence for the label distribution over manually labeled training data compared to system predictions. This model was chosen to label all 129,504 Regests.

For the project *Regesta Imperii* and Digital Humanities in general, our work offers the possibility to trace demographic groups (players) and themes through almost one thousand years of medieval history across different European locations. We showcased data analyses and visualizations. Manifold other possibilities may be explored in future work.

The Regesta Imperii in our opinion is a most challenging and linguistically interesting corpus. For historians, the RI is important as a fundamental source for medieval European studies. For linguists the RI may be very interesting due to its linguistic "uniqueness": syntactic constructions range from simple to most complex, the languages range from more modern German to different forms of medieval German to Latin. Great varieties in word forms exist. Semantically, the referenced objects and concepts are often confined to short periods of time. Thus, the RI presents challenges for researchers from many research fields. The challenging language, the considerable amount of data and the many interesting questions of humanities regarding the medieval times of Europe make the RI a great corpus for NLP researchers with special interest in Humanities.

Acknowledgments

We are grateful to the anonymous reviewers for their valuable comments.

References

Olivier Chapelle, Vikas Sindhwani, and Sathiya S. Keerthi. 2008. Optimization Techniques for Semi-Supervised Support Vector Machines. *J. Mach. Learn. Res.*, 9:203–233, June.

Morten W. Fagerland, Stian Lydersen, and Petter Laake. 2013. The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13(1):1–8.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *CoRR*, abs/1408.5882.

Andreas Kuczera. 2015. Graphdatenbanken für Historiker. Netzwerke in den Registern der Regesten Kaiser Friedrichs III. mit neo4j und Gephi. *Mittelalter. Interdisziplinre Forschung und Rezeptionsgeschichte*.

Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *ICML*, volume 32 of *JMLR Proceedings*, pages 1188–1196. JMLR.org.

D. Massad, E. Omodei, C. Strohecker, Y. Xu, J. Garland, M. Zhang, and L.F. Seoane. 2013. Unfolding History: Classification and analysis of written history as a complex system. Complex Systems Summer School Proceedings, Santa Fe Institute.

Albert Meroño Peñuela, Ashkan Ashkpour, Marieke van Erp, Kees Mandemakers, Leen Breure, Andrea Scharnhorst, Stefan Schlobach, and Frank van Harmelen. 2014. Semantic Technologies for Historical Research: A Survey. *Semantic Web Journal*.

Jan Paul Niederkorn. 2005. Julius von Ficker und die Fortführung der Regesta Imperii vom Tod Böhmers (1863) bis zu ihrer Übernahme durch die Kaiserliche Akademie der Wissenschaften in Wien (1906). In *Wege zur Urkunde, Wege der Urkunde, Wege der Forschung. Beiträge zur europäischen Diplomatik des Mittelalters (= Forschungen zur Kaiser- und Papstgeschichte des Mittelalters. Volume 24)*, Forschungen zur Kaiser- und Papstgeschichte des Mittelalters., pages 293–302, Köln, Weimar, Wien. Böhlau.

Michael Piotrowski. 2012. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Dieter Rübsamen and Andreas Kuczera. 2006. Verborgen, vergessen, verloren? Perspektiven der Quellenerschlieung durch die digitalen Regesta Imperii. In *Forschung in der digitalen Welt. Sicherung, Erschließung und Aufbereitung von Wissensbeständen. Tagung des Staatsarchivs Hamburg und des Zentrums Geisteswissenschaften in der digitalen Welt an der Universitt Hamburg am 10. und 11. April 2006*, Forschungen zur Kaiser- und Papstgeschichte des Mittelalters, pages 109–123, Hamburg. Rainer Hering, Jürgen Sarnowsky, Christoph Schäfer und Udo Schäfer.

Tuukka Ruotsalo, Lora Aroyo, and Guus Schreiber. 2009. Knowledge-Based Linguistic Annotation of Digital Cultural Heritage Collections. *IEEE Intelligent Systems*, 24(2):64–75.

Vikas Sindhwani and S. Sathiya Keerthi. 2006. Large Scale Semi-supervised Linear SVMs. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 477–484, New York, NY, USA. ACM.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. *CoRR*, abs/1509.01626.

Harald Zimmermann. 2000. *Die Regesta imperii im Fortschreiten und Fortschritt*, volume 20 of *Forschungen zur Kaiser- und Papstgeschichte des Mittelalters.* Böhlau, Köln, Weimar, Wien.