Classification of Medieval Documents: Determining the Issuer, Place of Issue, and Decade for Old Swedish Charters

Mats Dahllöf[0000−0002−4990−7880]

Uppsala University, Uppsala, Sweden
mats.dahllof@lingfil.uu.se

Abstract. The present study is a comparative exploration of different classification tasks for Swedish medieval charters (transcriptions from the "SDHK" collection) and different classifier setups. In particular, we explored the identification of the issuer, place of issue, and decade of production. The experiments used features based on lowercased words and character 3- and 4-grams. We evaluated the performance of two learning algorithms: linear discriminant analysis and decision trees. For evaluation, five-fold cross-validation was performed. We report accuracy and macro-averaged F1 score. The validation made use of six labeled subsets of SDHK combining the three tasks with Old Swedish and Latin. Issuer identification for the Latin dataset (595 charters from 12 issuers) reached the highest scores, above 0.9, for the decision tree classifier using word features. The best accuracy for Old Swedish issuer identification was 0.81. Place and decade identification produced lower performance scores for both languages. Which classifier design is the best one seems to depend on peculiarities of the dataset and the classification task. The present study does however support the idea that text classification is useful also for medieval documents characterized by extreme spelling variation.

Keywords: Text Classification, Medieval Charters, Old Swedish, Latin, Author Attribution, Automatic Dating.

1 Introduction

The present study is concerned with retrieval of metadata for medieval charters in Old Swedish and Latin by means of automatic analysis. The data source is the Diplomatarium Suecanum (SDHK), compiled by the Swedish National Archives. In particular, we focused on three tasks: the identification of the issuer of a charter, the place were it was issued, and the time of its creation. The issuer of a charter is the person in whose name the document is issued. In some cases an issuer is also the text author, as well as the material scribe, of the charter. In other cases, we can assume that authors and scribes were people employed by the issuer (e.g. a king) (see Wiktorsson's [9] investigation of Swedish medieval material). The issuer was the important person, whose authority gave validity to a charter. Methodologically, we can look upon the three tasks as typical scenarios for supervised text classification. This means that we explore the ability of machine learning models trained on labeled data to predict the relevant categories for

documents they have not seen before. Our approach to dating was a bin-based one, viz. prediction of the decade of production.

We performed a comparative exploration of the different classification tasks and different classifier setups. The methods we employed are all based on off-the-shelf implementations of algorithms and basic feature engineering. After the validation, we checked whether the conclusions we can draw for Old Swedish also are justified for the same kinds of charter in Latin. We assumed that the classification problems are roughly at the same level of difficulty for the two languages. However, we also expected to see some differences. The SDHK charters in Old Swedish exhibit an extreme variation in spelling [2]. They also use different inventories of letters. Furthermore, Old Swedish was in a state of transition from a more complex to a simpler nominal morphology during this period. Medieval Latin, by contrast, can be regarded as a more stable and standardized language, and it probably exhibits less variation in the charters. For that reason, we might expect that word form features are more useful for Latin, whereas Old Swedish word forms appear in many different spellings, thus weakening the signal they provide. Examples of alphabetic, spelling-related, and nominal case variation can be seen in the following instances of a conventionalized opening formula (All them/the men [who] see or hear this letter greet I/we. . .).

Alla dhe män detta breff hora ok see helsar iak (SDHK 7637, 1360)
Allom thøm thetta breff høra eller see helsar (SDHK 7846, 1360)
Allom theem thetta breef høra ællær see (SDHK 7729, 1360)
Alla the mæn thætta breef høra ælla see helsa iak (SDHK 11581, 1380)
Alle thee manne theta breff see eller höra helssr iagh (SDHK 11571,1380)
Alla the mæn thætta breff høræ ælla sea helsom wi (SDHK 11778, 1380)

2 Previous Work

Text classification is an important field of engineering in language technology (and one whose societal consequences at present are far-reaching). It is also a tool of crucial importance for digital humanities (DH). A challenging task in DH is the analysis of historical text. Pettersson [6] highlights seven aspects of historical documents which make them more difficult for natural language processing (NLP): spelling, vocabulary, semantics, morphology, syntax, sentence boundaries and sentence length, and code-switching. Another obstacle associated with older linguistic data is their scarcity, and the fact that diachronic change makes data from one period different from the material we may find from other periods. Another difficulty is that a few centuries ago written language was not at all standardized in the way we expect more recent writing to be.

The medieval stages of most languages, Latin possibly being an exception, are low-resource from an NLP point of view. There are a few examples of research on medieval North Germanic languages. Wahlberg, Mårtensson, and Brun [8] proposed the use of a statistical model for continuous dating of SDHK charters, working on a corpus of 5300 charters both in Latin and Old Swedish. They explored features deriving from both images and transcriptions, achieving a median absolute error of 12 years in the dating. Character 1- to 3-grams were used as transcription features. Boldsen and Paggio [1] also

investigated automatic dating of medieval charters, but in Danish, the closest relative of Swedish. Their corpus was fairly small (471 documents), deriving from a single convent archive. The charters were available in two levels of transcription, facsimile and diplomatic. They framed dating as a matter of classifying charters into bins[1] of 50-year periods, applying support vector machines (SVM) as classifiers. Word unigrams and character 1- to 3-grams proved most useful as features. Boldsen and Paggio [1] also saw that the information provided by the facsimile transcription, which, unlike the diplomatic one, preserves "palaeographic characteristics", was beneficial for dating.

Another study on the SDHK charters is the work of Karsvall and Borin [4] on named entity recognition for names of persons and places.

3 Diplomatarium Suecanum (SDHK)

The Swedish National Archives runs a long-term project whose aim is to make editions (and images) of the Swedish medieval charters available. In recent years the resource, the Diplomatarium Suecanum Main Catalogue (Svenskt Diplomatariums huvudkartotek, SDHK), has been put online.[2] The main series covers the time until 1380 and a second one 1401–1420. The charter records include regests (abstracts) and other kinds of metadata. For the most completely preserved and documented items, both facsimile images and the transcribed text have been published. In many cases only ancient copies, translations, or regests have been preserved. As the project is an ongoing one, many items are missing or incomplete, also besides the obvious lack of material for the periods 1381–1400 and after 1421. The languages of the transcribed charters are recorded in Table 1. The probably oldest preserved charter in Swedish in the National Archives is from 1344 (SDHK 5026). Latin was the main charter language in the earlier medieval period. As Sweden was a catholic country until 1527, Latin was also the language for all matters ecclesiastical to the end of the medieval period.

4 Experimental Setup

The three classification tasks, identification of issuer, place of issue, and decade of production, were approached in a uniform way. For each task, a subset of SDHK were used as labeled data (see Table 2). Due to the availability of charters with the relevant kinds of metadata specified, the sizes of the datasets and of the classes vary considerably. The training and evaluation of the dataset and classification design pairings were based on five-fold cross-validation.

4.1 Classifier Designs

The experiments explored features based on lowercased words and character 3- and 4-grams. The n-grams were produced from words with B and E added at the beginning

[1] Bin-based approaches to dating (like that of Boldsen and Paggio [1] and the present study) turns the problem of dating into one of using labels belonging to a nominal scale when time really is a matter of an interval scale.

[2] `https://sok.riksarkivet.se/SDHK`.

Table 1. The languages represented in the SDHK charters for which transcriptions are available. Statistics compiled from the downloaded data. (A few obvious labels omitted in the electronic version were not corrected.)

Language	**Number of charters**
Latin	8212
Old Swedish	3791
Latin/Swedish mixtures	15
Medieval translations into Swedish	27
German	410
Norwegian	102
Danish (including translations and uncertain items)	29
Dutch, French, and Italian	Few

and the end, respectively. So, for instance, a word like *Swea* generates these 3-grams: *Bsw*, *swe*, *wea*, *eaE*. As weights for the features, we used their relative frequencies, standardized as described below. Relative frequency is a quantity that "neutralizes" the length of documents, and has for a long time been used in e.g. authorship attribution [3]. Tokens containing Arabic numbers were removed (on roman numerals, see below). We selected features based on their ranking in terms of descending mean (over charters) relative frequency (again in order to "neutralize" the length of documents). The documents were modeled by selections of $n \in \{500, 1000, 2000, 3000, 4000\}$ top-ranking features (or all if the size of the full set was smaller than that).

In order to prevent the classifiers from possibly learning to "read" the classes directly from explicit expressions in the charters, we removed all tokens which are similar to the tokens in the list of given classes. The similarity was operationalized in terms of relative (divided by the length of the longest of the two compared strings) Levenshtein distance, with 0.25 as the threshold. For instance, among the tokens that are removed because they are too similar to issuer names in Table 2, we find, for instance, *albrect*, *albrikt*, *tordh*, *bondæ*, *bonde*, *erikz*, *erikx*, *ärlansson*, *niclesson*, *niclæsson*, *niclisson*, and *sthen*. This threshold seemed to represent a reasonable sensitivity for this purpose. For the decade bin task, we tried to filter out roman numerals, which often stand for years (as did [8], automatically, and [1], manually). There are often spaces and points in the year numerals, e.g. "M. CCC. XXX. IX" (SDHK 4545). So, we removed all tokens (which are lowercased) containing one of the following substrings: *ccc*, *iv*, *xc*, *xl*, *ix*, *xi*, *xx*, *vii*. Again, this is a heuristic solution which we assume achieves a high recall for tokens referring to years. Obviously, also other tokens, e.g. numerals serving other semantic purposes, are removed to some extent.

Our experiments employed two classifier learning algorithms, implemented in the *scikit-learn* library of "simple and efficient tools for data mining and data analysis" [7]. They had performed well and efficiently in preliminary studies.

- Decision Tree Classifier: A "non-parametric supervised learning method" which creates "a model that predicts the value of a target variable by learning simple decision rules inferred from the data features."[3]
- Linear Discriminant Analysis (LDA): "A classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule."[4]

We standardized the training feature vectors by removing the mean from the feature values and scaling them to unit variance and then scaling the unseen vectors in the same way during prediction.[5]

4.2 Evaluation and Performance Metrics

The evaluation of the classifiers took place by means of a cross-validation procedure. This means that the data are partitioned in a certain number (here five) of subsets (folds) and that the same number of models are trained, each time reserving one fold (unseen in the training process) for evaluation. Each element of the data afterwards corresponds to one evaluation prediction. The datasets were randomly partitioned into stratified folds, i.e. the proportions of the various classes stayed the same in the folds as in the whole set. We recorded accuracy, i.e. the fraction of predictions which are correct, and the macro-averaged F1 score[6] (mF1) which gives equal weight to the different classes [5]. For the "plain" accuracy score each instance of large classes counts as much as those from smaller ones. We first analysed the results for the Old Swedish tasks. As a second step, we used the charters in Latin as a test set, and examined to what extent the conclusions drawn from Old Swedish also hold for the Latin data.

5 SDHK Datasets for the Three Prediction Tasks

The editions from which the SDHK transcriptions derive have been published in a couple of different series since 1829 and the work is still ongoing. As can be expected, the transcriptions have been produced according to changing scholarly standards. The current editor-in-chief [2], characterizes the transcription level, "modified diplomatic", as one that allows "cautious normalization" (our translations).

We compiled datasets for the various classification tasks based on the availability of labeled charters in SDHK.[7] The transcriptions of the charters in the datasets have a length of at least 40 tokens (continuous letter sequences). For each task, a subset of

[3] `sklearn.tree.DecisionTreeClassifier` [7].

[4] `sklearn.discriminant_analysis.LinearDiscriminantAnalysis` [7].

[5] Using `sklearn.preprocessing.StandardScaler` [7].

[6] Plain F1 is the harmonic mean of the precision and recall.

[7] The full set of SDHK charters was extracted from the HTML files of the National Archive website, `https://sok.riksarkivet.se/SDHK`, as downloaded on October 5, 2019. We removed HTML tags from the transcriptions. The cleaned charter data (an XML file) and Python code for performing the experiments reported here are available open access as supplementary material at `http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-400834`. (We also provide code for downloading SDHK charters and preprocessing them.)

Table 2. The three labeled Swedish SDHK datasets. The units are transcribed charters. The issuer and place names are substrings from the metadata, except the two "Kung Erik" (King Eric), which were disambiguated from the date. In some cases, co-issuers have been disregarded, as has additional information about the place of issue, e.g. "Vadstena kloster" (abbey) is labeled with the same label as a plain "Vadstena". A decade like "d1410" is the period 1410–1419. The 1390s are missing due to the lack of transcriptions.

Issuers (12)	(486)
Kung Erik (of Pomerania)	80
Kung Magnus	70
Bo Jonsson	59
Sten Bosson	56
Ivar Nilsson	53
Tord Bonde	46
Kung Albrekt	45
Sten Bengtsson	28
Kung Erik (Magnusson)	14
Kung Håkan	13
Elof Djäken	12
Nils Erlandsson	10

Places (23)	(1516)
Vadstena	248
Uppsala	171
Stockholm	147
Västerås	109
Linköping	108
Strängnäs	107
Skara	77
Skänninge	77
Söderköping	66
Åbo	63
Nyköping	57
Nydala	39
Kalmar	38
Västervik	27
Helsingborg	25
Lund	25
Jönköping	22
Lödöse	22
Arboga	20
Enköping	19
Växjö	17
Örebro	17
Marieborg	15

Decades (8)	(3715)
d1410	1108
d1400	932
d1370	702
d1360	426
d1350	285
d1380	110
d1420	108
d1340	44

Table 3. The datasets for SDHK charters in Latin with the same number of classes as those for Old Swedish, see Table 2, and a roughly similar distribution of class sizes. (*Biskop*: bishop, *ärkebiskop*: archbishop, *hertigarna*: dukes, *kardinalpräst*: cardinal priest.)

Issuers (12)	(595)
Kung Valdemar	99
Kung Albrekt	86
Julianus biskop i Bertinoro	84
Kung Birger	81
Ärkebiskop Nils	71
Styrbjörn i Strängnäs	33
hertigarna Erik och Valdemar	33
Ärkebiskop Peter i Lund	31
Dominicus kardinalpräst	24
Antonius biskop i Luni	23
Biskop Peter i Linköping	20
Philippus kardinalpräst	10

Places (23)	(1913)
Uppsala	246
Linköping	219
Lund	218
Lübeck	195
Villeneuve	123
Skänninge	83
Strängnäs	81
Arnö	64
Kalmar	64
Helsingborg	58
Skara	58
Stralsund	57
Söderköping	57
Paris	56
Lödöse	44
Nyköping	43
Sigtuna	43
Visby	42
Varberg	38
Viterbo	37
Rostock	32
Anagni	29
Lyon	26

Decades (8)	(4904)
d1340	998
d1350	988
d1360	884
d1330	680
d1370	591
d1400	404
d1410	300
d1380	59

SDHK was selected as labeled data (Table 2). We did this by creating lists of labels which gave us a reasonable number of instances for the smallest classes. There is also a requirement that they be explicitly dated between 1250 and 1500. Otherwise, all charters matching the labels are included. The number of charters in each dataset and for each class varies. As can be expected, issuers are associated with the smallest number of charters, whereas larger lots originate from identifiable places and decades. So, we have 12 issuers of 486 charters, 23 places for 1516 charters, and 8 decades of issue (the 1390s being a gap) for almost 5000 charters. We also compiled a dataset (Table 3), to be used for testing, with charters in Latin with the same number of classes as those for Old Swedish and a roughly similar distribution of class sizes.

6 Results

6.1 Classification of Old Swedish Charters

The performance of the classifier setups based on the two learning algorithms, the various kinds of feature, and the different (maximal) numbers of features when applied to the three Old Swedish datasets is plotted in Figure 1. We see that the relative performances of the setups are quite different for the different classification tasks. The two scores, accuracy and mF1, in most cases support the same conclusions about the relative merits.

For issuer prediction, the highest accuracy (0.81) occurred with LDA and 2000 3-grams, but the highest mF1 (0.73) with 500 word features. The decision tree algorithm clearly showed a weaker performance on this dataset. We also see how the curve for LDA with word features dives as their number increases.

A quite different situation obtained for the place of issue classification. Here, the LDA algorithm combined with the largest number (4000) of 4-grams gave the best performance for both scores (accuracy=0.72, mF1=0.60). The second best option was using 3-grams with LDA, but their usefulness seemed to peak when we take about 2000 of them.

For the decade bin dating, the decision tree classifier with the largest selection of 3-grams clearly outperformed all other setups (accuracy=0.76, mF1=0.70).

6.2 Testing the Classifier Setups on Charters in Latin

When we performed the same series of experiments on the Latin datasets (Table 3), we got the scores plotted in Figure 2. Issuer identification yielded an outcome that is clearly different from what we saw for the Old Swedish data. In this case, the decision tree algorithm combined with word features gave the best performance for 500 and 4000 features (no difference at two decimals: accuracy=0.91, mF1=0.90). The scores are considerably higher than for Old Swedish. (Again, we see how the LDA curve dives for word and 3-gram features, but not for 4-grams.)

For place of issue classification, the LDA algorithm combined with 3000 3-grams gave the best performance as reflected by both scores (accuracy=0.71, mF1=0.63). The second best option was the same setup, but with 2000 features. For this task, the scores

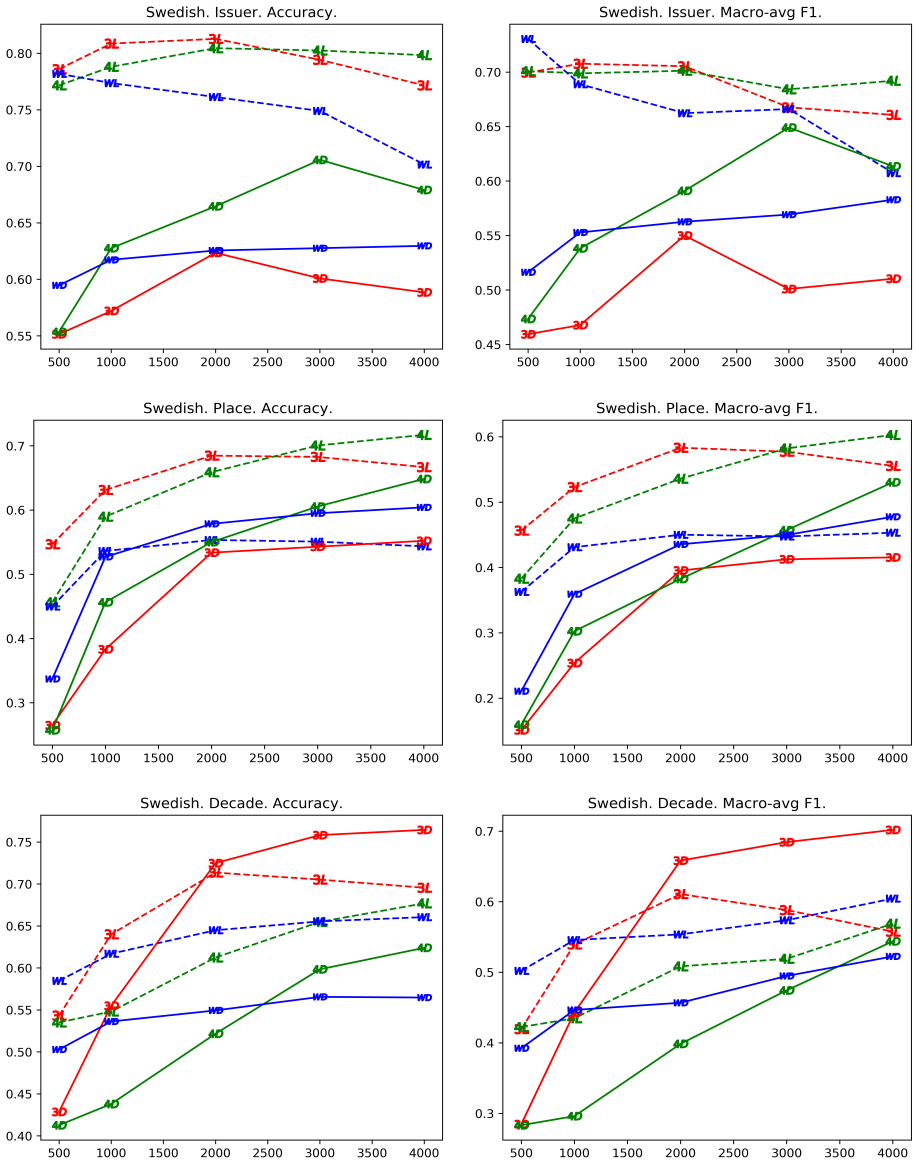

Fig. 1. The performance of the classifiers for the Old Swedish charters by accuracy (left) and macro-averaged F1 score (right). (Compare Figure 2.) *w*: graph word. *3* and *4*: for 3- and 4-grams, respectively. *D*: Decision Tree, *L*: Linear Discriminant Analysis. Number of features on the x-axis and scores on the y-axis.

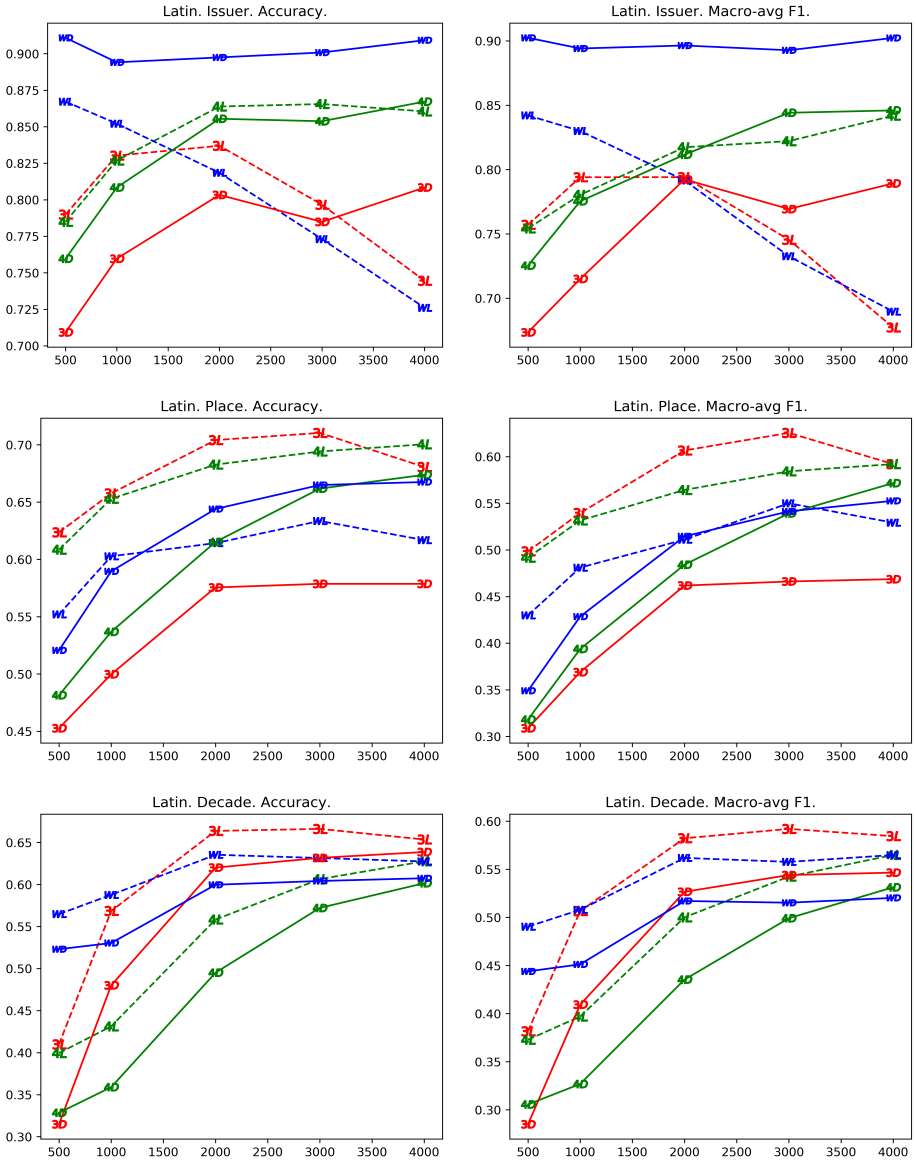

Fig. 2. The performance of the classifiers for the Latin charters by accuracy (left) and macro-averaged F1 score (right). (Compare Figure 1.) *w*: graph word. *3* and *4*: for 3- and 4-grams, respectively. *D*: Decision Tree, *L*: Linear Discriminant Analysis. Number of features on the x-axis and scores on the y-axis.

for the Latin and Old Swedish datasets were quite close to each other. For decade classification in Latin, the two best scoring setups were the same as for place prediction: LDA, with 3000 3-grams performed the best (accuracy=0.67, mF1=0.59), with the 2000 3-gram setup just a little lower. Interestingly, this was the only task for which the Latin data presented lower scores than those we saw for Old Swedish.

7 Discussion and Conclusions

For both Old Swedish and Latin, issuer identification is the task that gives the highest scores. A possible and partial explanation is that individual issuers are associated with particular scribes, circumstances, and functions, as well as specific places and often short periods. (The number of classes for the various tasks is, of course, also important.) Places and decades, we can assume, are connected to more heterogeneous sets of charters. For both languages, we see how the decision tree algorithm makes successful use of an increasing number of word features, whereas the rarer words hurt the performance of LDA. Nevertheless, LDA in other setups performs best for Old Swedish issuer prediction.

We expected that words (as spelled) in Latin would be more useful as features than for Old Swedish, where word forms appear with many different spellings, and consequently with lower frequency. This was however only confirmed in the issuer identification scenarios. In cases of high degrees of spelling variation, character *n*-grams have the ability both to indicate the presence of lexemes and to capture idiosyncratic properties of spelling.

Another interesting similarity between Old Swedish and Latin is seen in the way the LDA 4-gram curve is rising with an increasing number of features for the task of of place of issue prediction. On the other hand, the rarest 3-gram features seem to do more harm than good.

When it came to decade bin dating for Old Swedish, all 4000 3-grams proved helpful when fed to the decision tree algorithm. (In the Latin case, 4-grams were in a similar way helpful for the algorithm, even if other setups were more successful.)

We could also see that Old Swedish was easier to date as far as the classifiers we tried go. This might reflect the diachronic changes of various aspects of Swedish during the medieval period. Latin had presumably a more conservative and standardized appearance, which might explain why the Latin charters turned out to be somewhat more difficult to date by the methods we explored.

The present study has shown that text classification applied to “noisy” medieval data has the ability to perform on a level which is arguably useful as a support for philologists or historians aiming for a closer analysis. A methodological lesson is that the relative performance of classifier setups is very sensitive to the properties of the dataset to which they are applied. We cannot say that there is a particular approach that should be considered the generally preferable one for Old Swedish or medieval charters. Rather, which classifier design is the best one can only be determined if we take the specific kind of classification task and the peculiarities of the dataset into consideration.

Acknowledgements

This work has been carried out in the project *New Eyes on Sweden's Medieval Scribes. Scribal Attribution using Digital Palaeography in the Medieval Gothic Script* led by Lasse Mårtensson and funded by Riksbankens Jubileumsfond (Dnr NHS14-2068:1).

References

1. Boldsen, S., Paggio, P.: Automatic Dating of Medieval Charters from Denmark. In: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, CEUR Workshop Proceedings. Vol. 2364, pp. 58–72 (2019).
2. Gejrot, C.: Medeltiden i dataåldern om svenskt diplomatarium idag. In: Kungl. Vitterhets historie och antikvitets akademien årsbok. 2011, pp. 97–108 (2011).
3. Holmes, D.I.: Authorship Attribution. Computers and the Humanities, 28, pp. 87–106 (1994).
4. Karsvall, O., Borin, L.: SDHK meets NER: Linking Place Names with Medieval Charters and Historical Maps. In: Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference, CEUR Workshop Proceedings. Vol. 2084, pp. 38–50 (2018).
5. Kestemont, M., Tschugnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In: Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings. Vol. 2125 (2018).
6. Pettersson, E.: Spelling Normalisation and Linguistic Analysis of Historical Text for Information Extraction, Uppsala: Acta Universitatis Upsaliensis (2016).
7. scikit-learn Machine Learning in Python homepage, "Multiclass and multilabel algorithms," `https://scikit-learn.org/stable/modules/multiclass.html`, last accessed 2019/09/23.
8. Wahlberg, F., Mårtensson, L., Brun, A.: Large scale continuous dating of medieval scribes using a combined image and language model. In: 12th IAPR Workshop on Document Analysis Systems (DAS 2016), pp. 48–53 (2016).
9. Wiktorsson, P.-A.: Skrivare i det medeltida Sverige I. Skara: Skara stiftshistoriska sällskap (2015).