Automatic Dating of Medieval Charters from Denmark*

Sidsel Boldsen[1 [0000−0002−6880−5345]] and Patrizia Paggio[1,2 [0000−0002−2484−2275]]

[1] University of Copenhagen
{sbol,paggio}@hum.ku.dk
[2] University of Malta
patrizia.paggio@um.edu.mt

Abstract. Dating of medieval text sources is a central task common to the field of manuscript studies. It is a difficult process requiring expert philological and historical knowledge. We investigate the issue of automatic dating of a collection of about 300 charters from medieval Denmark, in particular how n-gram models based on different transcription levels of the charters can be used to assign the manuscripts to a specific temporal interval. We frame the problem as a classification task by dividing the period into bins of 50 years and using these as classes in a supervised learning setting to develop SVM classifiers. We show that the more detailed facsimile transcription, which captures palaeographic characteristics of a text, provides better results than the diplomatic level, where such distinctions are normalised. Furthermore, both character and word n-grams show promising results, the highest accuracy reaching 74.96 %. This level of classification accuracy corresponds to being able to date almost 75 % of the charters with a 25-year error margin, which philologists use as a standard of the precision with which medieval texts can be dated manually.

Keywords: Automatic dating · Medieval charters · Language models

1 Introduction

Dating of medieval text sources is a central task common to the field of manuscript studies. The majority of medieval manuscripts are without explicit reference to the time and place they were produced and by whom they were written. This knowledge, however, is crucial in order to interpret the content and context of a source. For example, philological research on historical text is very dependent on correct interpretation of word forms, which is only possible when knowing the

* This study was conducted at the University of Copenhagen within the project *Script and Text in Time and Space*, a core group project supported by the Velux Foundations. A general description of the project is available from https://humanities.ku.dk/research/digital-humanities/projects/writing-and-texts-in-time-and-space/. We thank the project, in particular Alex Speed Kjeldsen, for making the data available to this study.

CEUR-WS.org/Vol-2364/5_paper.pdf

origin of the given source. The dating of medieval texts is often a long and laborious process requiring expert philological and historical knowledge. Introducing automatic methods to facilitate this process, therefore, is a valuable effort.

Dating may rely on a range of different criteria, including characteristics of the handwriting in a document, the material state of the parchment or paper, reference to historical events in the manuscript, linguistic evidence, etc. Generally, precise dating is very difficult to achieve, and an error rate of 25 years is considered acceptable.

In this paper, we investigate the issue of automatic dating of charters from medieval Denmark, in particular how n-gram models based on different transcription levels of the charters can be used to assign the manuscripts to a specific temporal interval.

While seeking to develop knowledge on how far Natural Language Processing (NLP) methods can take us in attacking the problem of medieval manuscript dating, we also want to determine how different levels of transcription produced according to recommended philological standards contribute to this task. In particular, we will look at two levels of transcription, namely (i) a facsimile transcription in which variations in handwriting are represented, and (ii) a diplomatic transcription in which such differences are normalised, but where differences in spelling are still present. To the best of our knowledge, this is the first attempt at capitalising on the use of different philological transcription levels for automatic dating of documents.

2 Background

Previous attempts at automatic dating of medieval charters fall into two basic groups depending on whether they use visual features of the printed materials obtained through image processing, or whether they use language models. In a few cases, a combination of visual and language features is used. An additional and orthogonal distinction concerns whether the task is approached as continuous dating along the timeline or classification into a number of time intervals.

Visual features capturing the strokes of handwritten characters were used for instance in [5] and [6] for the automatic dating of a collection of 1,706 medieval documents from the Dutch language area. Dating was treated as a regression problem in the former study, and as classification in 25-year intervals in the latter, which reports a mean absolute error of 20.9 years for the whole dataset.

In [16] and [17] visual features were extracted to train various models, in particular regression models as well as Convolutional Neural Networks (CNN) for continuous dating of medieval charters from the *Svenskt Diplomatariums Huvudkartotek* (SDHK) collection, which contains over 11,000 charters in Latin and Swedish, of which about 5,300 are transcribed. These studies report absolute errors of 18.3 and 36.8 years at the 50th and 75th percentiles, respectively, for a Support Vector Regression (SVR) model. This corresponds to classifying 50 % of the dataset with an error of ±18.3 years and 75 % with an error of ±36.8

years. For a CNN model, the absolute error is of 10 and 22 years at the 50th and 75th percentiles, respectively.

In [15], visual features were combined with language models to train several Gaussian Mixture Models (GMM) for the same regression task. While the visual features model the changes in pen stroke over the years, the language features are character n-grams aiming at capturing changes over time of short character sequences. The combined image and language model performs with an absolute error of 12 years for 50 %, and 22 years for 75 % of the dataset, and constitutes an improvement compared to similar GMM models only trained on visual features.

In NLP research, dating of documents is usually approached as a temporal document classification task. In contrast with the studies mentioned above, visual features extracted from physical texts are ignored and instead the various approaches try to capitalise on the way the lexicon, the morphology or the syntax of a language changes over the years. The evaluation measures reported in NLP classification studies do not generally refer to error measures, but rather to precision or accuracy relative to different granularities of the temporal intervals (or bins) used, and compared to a more or less naive baseline.

An example of methods based on lexical knowledge is presented in [1], where temporal text classification is based on change in term usage, while [2] used the Google Books Ngram corpus to identify neologisms and archaisms for the dating of French journalistic texts. Similarly, in [11] the same lexical resource was used to assign political terms to temporal epochs of varying length depending on their usage change. Stylistic features such as average sentence and word length, lexical density, and lexical richness were used in [13] for the temporal classification of Portuguese historical texts.

A diachronic text evaluation task [12] was proposed as part of the SemEval 2015 initiative. The task consisted in the temporal classification of newspaper text snippets from 1700 to 2010 into time intervals of different sizes. The best model was a multiclass Support Vector Machine (SVM) classifier using stylistic features such as character, word and part of speech (POS) tag n-grams, but also external estimates from the Google syntactic n-gram database, and achieved an accuracy of 54.3 % on the 20-year interval classification task [14]. Word n-grams of order 1–3 with and without their POS tags were also used in [18] to train models for the temporal classification of Portuguese historical documents from the period 16^{th} to early 20^{th} century. The study reports an accuracy of 74.1 % for the best SVM classifier obtained in the task of temporal classification in 100-year temporal bins.

Character n-grams were used in [10] to calculate the distance between historical varieties of Portuguese. The authors argue that character sequences capture not only morphological and lexical, but also phonological differences between language varieties. Interestingly for our purposes, the study experiments with two different styles of transcription of the original texts, and shows that the best results in distinguishing historical variants are obtained with the transcriptions that preserve the original spelling instead of standardising it.

To sum up, language features, in particular word and character n-grams, have been applied with a certain degree of success to temporal text classification of relatively modern text collections and to quantify the difference between historical texts of different periods, but not to any large extent to the specific task of medieval manuscript dating. Some evidence, however, has been presented that they may contribute a useful addition to image-based features for that task. Our goal in this paper is to provide additional evidence in this direction.

3 The charters of St. Clara Convent

The study revolves around the collection of charters that belonged to St. Clara Convent outside the city of Roskilde in Denmark. The charters document the property and status of the convent and they date from when it was founded in 1256 till it was closed after the Reformation. In 1561 the properties and buildings of St. Clara became part of the University of Copenhagen and so did its archives. The collection of charters is now part of the Arnamagnæan Collection [3].

The St. Clara Convent archive contains 471 charters in total. They are written in several languages, most of them in Latin and Danish, and a few in Swedish and Low German. Most of the charters are originals, handwritten on parchment and with a seal attached, while others are copies of original charters and are handwritten on paper.

Most of the original charters can be time-stamped, either from explicit dates in the text or indirectly from knowledge about the scribe and the persons mentioned. The copies are more difficult to date: They do not have an explicit time stamp from when they were written and since the content is a copy of earlier text, the historical context cannot provide the dating of the charter either. Instead knowledge about spelling variation and palaeographic differences, historical linguistics, or material evidence about paper and ink, may be used to assign a possible date [7].

The charters of the collection are being prepared for a digital scholarly edition. First, digital photographs of the handwritten manuscripts are produced. Then, the charters are transcribed at different levels of detail, namely *facsimile*, *diplomatic*, and *normalised* levels. These three levels of transcription are recommended by Menota (Medieval Nordic Text Archive) [4] as means of encoding medieval manuscripts through text representations as close to the original manuscript as possible. In the facsimile transcription, palaeographic characteristics, in other words differences due to handwriting, are encoded. For example, different ways of writing a 'd' are represented by distinct characters (e.g., 'd' or 'ꝺ') and different types of abbreviating diacritics are preserved. In the diplomatic transcription such differences are normalised so that characters that are not phonologically contrastive will be unified to a single character at this level of transcription. Furthermore, all abbreviating symbols are expanded. Finally, at the normalised level spelling variation is reduced to a common standard. In addition to the three levels of transcription, all text in the charters will be lemmatised.

(a) (b)

Fig. 1: AM Dipl. Dan. Fasc. LI 3. Founding letter of the convent. The first will of Ingerd af Regenstein witnessed by the bishop of Lund in 1256, where she declares her intention of founding the convent.

To illustrate the differences between the different levels of transcription, consider the first line of the text in Figure 1(b):

(a) ſoꝛoꝛ ⁊ ꝯonaſteꝛíj earū̃ ıꝯ poſterum
(b) soror(um) (et) monasterij earu(m) in posterum
(c) sororum et monasterii earum in posterum

In the facsimile transcription (a) different abbreviations are annotated, e.g., the Tironian *et*, ⁊, which was used by scribes in the medieval period, and different allographs are represented by different characters, e.g., í vs. ı, and ꝯ vs. ꝯ. At the diplomatic level (b) the abbreviations have been expanded, marked by parentheses, and the allographic variation has been normalised and, thus, we only find one type of *i* and only one *m*. In the normalised transcription (c) spelling is standardised. In this particular example this only includes the spelling of final *i* as *j* in *monasterij*.

So far 293 charters have been transcribed at the facsimile level and a diplomatic transcription has been generated automatically. Out of these, 291 are originals and are dated either through explicit dating or based on the content of the text. Two of the charters are copies of original manuscripts and have not yet been dated. One of these originals is among the transcribed documents, while the other is not known. The 291 transcribed and dated charters will constitute the dataset of this article. Using both the facsimile and the autogenerated diplomatic level of the dated originals, we wish to test how these different levels of transcription (capturing spelling variation and palaeographic differences) can be used to model the production date of the medieval charters.

To give an idea of the difference between the facsimile and diplomatic transcriptions of the charters in terms of how much the variation is reduced across the two transcription levels, in Table 1 we report token counts related to the

Table 1: Token counts for charters in Latin and Danish taken from the facsimile (facs.) and diplomatic (dipl.) transcription levels.

Language	word tokens		char tokens	
	facs.	dipl.	facs.	dipl.
Latin	14,169	12,146	231	132
Danish	7,662	7,401	173	100

two levels for the Latin documents and those in Danish. From the counts it can be calculated that the reduction in word token counts is of 14 % for the Latin manuscripts and 3 % for the Danish ones, while it amounts to 43 % for Latin and 42 % for Danish when we look at character token counts.

4 Methodology

In this study, dating of documents is dealt with as a classification task in which the charters are classified as belonging to a time interval, or bin. Two sets of bins are considered: dividing the documents into (i) four classes of 100 years (corresponding to the 13^{th}, 14^{th}, 15^{th}, and the 16^{th} centuries), and (ii) eight classes of 50 years each (two classes pr. century, i.e., 1250-1300, 1300-1350, ..., 1550-1600). The division of the timeline into periods is naive in the sense that the boundaries are not based on any knowledge about linguistics or historical periods; it is simply a division of the timeline into series of bins of equal size. Furthermore, framing the problem as a classification task is a simplification, since it makes no assumptions about documents from two time spans close to each other being closer than two documents belonging to time spans further away. However, if accurate, such a method would still be useful in providing an approximate assessment of their possible date of production. For instance, classifying the documents in 50-year bins can be seen as a way to date the documents with a 25-year error margin by assigning all the documents belonging to one category the median year of the range. We also performed classification in 100-year bins, in spite of it being very coarse-grained, to position our work against results from the literature, in particular [18].

A number of classification experiments are reported here. In all of them, each of the charters is represented as a vector, the values of which correspond to the frequency in the charter of either word or character n-grams of order 1-3. Different experiments are run with n-grams extracted from the facsimile and diplomatic levels, and we also tried combining unigrams and bigrams with trigrams, again separately for the two transcription levels.

SVMs were used to classify the charters. First of all, SVMs are known to work well with sparse representations, which are a potential problem when using n-grams of a larger order together with a relatively small dataset. Secondly, when applied to document classification tasks such as the identification of similar languages (e.g., discriminating between Dutch and Flemish) this model provides

state-of-the-art results [9, 19]. The task of document dating is somewhat related to this task if one considers the stages of developments of a language to be similar to dialects or very closely related languages.

When carrying out the experiments, two baselines are considered. The first one always chooses the most frequent class, which has slightly different likelihoods of being correct depending on the size of the time spans. The second one picks the most frequent class for each of the languages. Here we chose to group the Swedish and Low German together with the Danish as one group. Since the two language groups considered (Latin and Danish, misc) are not equally distributed, the average accuracy of this baseline is a weighted mean of the accuracy that would be reached for each language separately. Again, two different measures are obtained depending on the intervals chosen. The reason why we chose to add the second baseline is due to the fact that the distribution of the documents over languages correlates with time (as we will see in the following section). Thus, in order to control that the model is not 'only' performing language classification, we use this baseline to test whether the models can outperform it. The reader should keep in mind that we do not provide explicit knowledge of the source language of a charter to the models that we train in the next section. That knowledge is only available implicitly through the text given as input.

In all the experiments 10-fold cross validation was used to evaluate the different models.

5 Dataset

As mentioned earlier, the dataset used in this study is constituted by the 291 charters from the collection that have been transcribed so far. Two different levels will be included: (i) the facsimile transcription, where allographic variation is annotated, and (ii) the diplomatic transcription, where it is normalised, while spelling variation is still maintained. The dataset contains documents in the four languages represented in the collection. The distribution of the charters amongst the different languages can be seen in Table 2.

Table 2: The 291 charter collection: language statistics.

Language	Counts	Proportion
Latin	209	0.72
Danish	78	0.27
Swedish	2	0.005
Low German	2	0.005
Total	291	1

In Figures 2 and 3 the charters are grouped into bins of 50 and 100 years, respectively. As can be seen, the Latin documents are most dominant in the 13^{th}

Table 3: The 291 charter collection: distribution over time.

Period	Counts	Proportion	Cumulative proportion
1250-1300	78	0.27	0.27
1300-1350	60	0.21	0.48
1350-1400	36	0.12	0.60
1400-1450	51	0.17	0.77
1450-1500	39	0.13	0.90
1500-1550	25	0.09	0.99
1550-1600	2	0.01	1.00

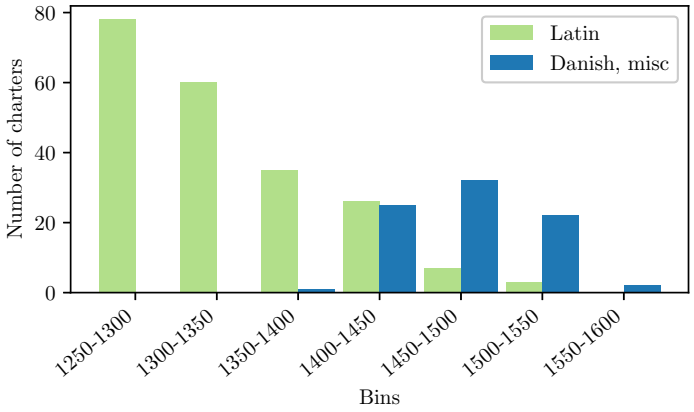


Fig. 2: Plot of the distribution of the charters in 50-year bins.

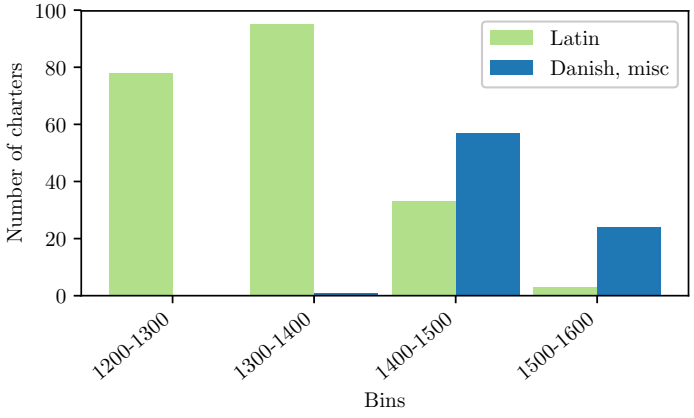


Fig. 3: Plot of the distribution of the charters in 100-year bins.

and 14th centuries, whereas there is a shift during the 15th century to documents being written in Danish. The two Low German documents are from 1350-1400 and from 1400-1450, while the two Swedish ones are from 1500-1550. Furthermore, since there are more Latin documents than Danish, the total distribution of documents over time is skewed such that documents from the middle of the 13th century to the end of the 14th century constitute almost 50 % of the documents in total (see the cumulative proportions in Table 3).

All the transcriptions were preprocessed by removing all dates and adding additional document start and end symbols to preserve this positional information when representing the documents as n-gram counts.

6 Results and discussion

Two baselines were chosen, as already mentioned, one always choosing the most frequent class, and the other also relying on knowledge of the language groups. Table 4 shows the accuracy and weighted F1 score reached by these baselines depending on the bin size. Whilst the accuracy is the proportion of correct predictions, the weighted F1 score is based on precision and recall, and is computed by taking the average F1 score of the predicted classes and multiplying it by the proportion of supporting instances. We chose this measure instead of a simple F1 score to account for the imbalance between the classes.

Table 4: Baseline accuracy and weighted F1 scores for two temporal bin sizes.

Baseline	Model	50-year bins		100-year bins	
		acc	F1	acc	F1
I	majority class	26.80	11.33	32.99	16.36
II	majority class for each language	37.80	36.84	52.23	50.75

A total of 32 different experiments were run, 16 for each transcription level. We trained models using unigrams, bigrams and trigrams, as well as a combination of unigrams, bigram and trigrams, built from characters and from words, and including labels for the two different time interval sizes. The results for accuracy are displayed in Tables 5 and 6, while the results for the weighted F1 scores are displayed in Tables 7 and 8. Tables 5 and 7 show the results obtained using the facsimile transcription and Tables 6 and 8 show those relating to the diplomatic level. The accuracy and weighted F1 scores correspond to the average scores obtained by the classifier over the 10 folds in each experiment.

In the tables, the highest accuracy and F1 score have been highlighted for the word and character models respectively. The models at the facsimile and diplomatic levels for the 50-year bins show the same pattern for both accuracy and F1 scores: In the character models, the scores *increase* when increasing the order of the n-gram. In word models, in contrast, the scores *decrease* when increasing the

Table 5: Accuracy scores using the facsimile transcription.

N-Gram	50-year bins		100-year bins	
	Char	Word	Char	Word
1	70.73	**74.96**	79.67	**79.63**
2	73.12	53.30	78.77	64.41
3	74.48	37.93	80.14	40.28
1-3	**74.90**	72.32	**80.88**	76.26

Table 6: Accuracy scores using the diplomatic transcription.

N-Gram	50-year bins		100-year bins	
	Char	Word	Char	Word
1	60.04	**68.83**	71.50	**74.72**
2	70.81	50.25	**79.19**	66.86
3	70.77	36.56	75.95	44.50
1-3	**73.27**	66.56	77.09	70.67

Table 7: F1 scores using the facsimile transcription.

N-Gram	50-year bins		100-year bins	
	Char	Word	Char	Word
1	69.73	**73.83**	79.40	**78.75**
2	71.87	48.21	78.28	60.17
3	73.45	25.55	79.26	27.58
1-3	**73.79**	70.70	**80.15**	74.13

Table 8: F1 scores using the diplomatic transcription.

N-Gram	50-year bins		100-year bins	
	Char	Word	Char	Word
1	58.58	**67.06**	71.16	**73.91**
2	69.70	43.52	**78.24**	62.86
3	68.98	24.53	75.02	34.59
1-3	**71.60**	64.40	75.78	68.38

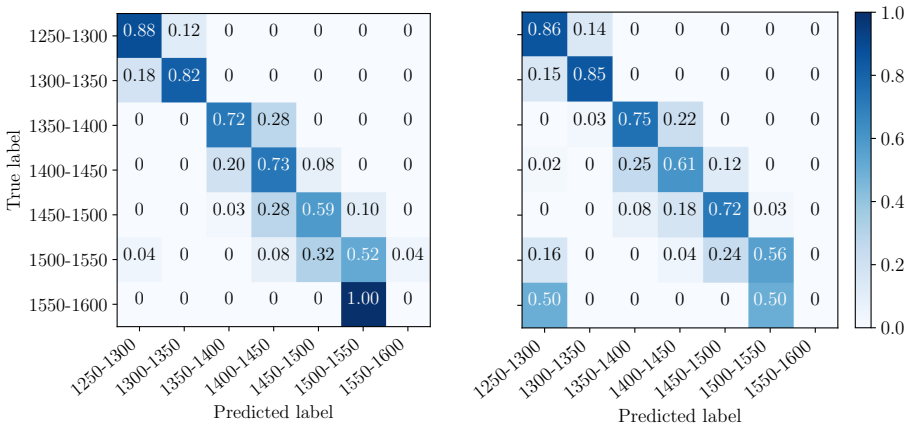
order of the n-gram. One possible explanation of why this happens is that the dimension of the vector representations built over word n-grams is too high, and therefore causes the models to overfit to the training data. Compared to the facsimile character model that has 269 unique unigrams and 29,387 trigrams (159 and 12,384, respectively, using the diplomatic transcription), the word model has 21,704 unique unigrams and 64,363 trigrams (19,595 and 59,400, respectively, using the diplomatic transcription). One possible way of circumventing this issue could be to perform some type of feature selection on the input features. In this way one would be able to reduce noise at the same time as reducing the high dimensionality of the input. Such a reduction in dimensionality would also limit the sparseness of the input space compared to the amount of data available to the models.

In general the models using the facsimile transcription exhibit higher accuracy and F1 scores than the corresponding models using the diplomatic transcription. When using only single character counts, for example, we reach an accuracy of 70 %, compared to 60 % with the same model using the diplomatic level. The fact that using facsimile transcription yields more accurate results than relying on the diplomatic transcription, confirms our expectations given the knowledge we have of the importance of palaeographic differences for the dating of medieval text. However, when increasing the order of the n-gram for the character models the difference becomes smaller. It would be interesting to compare what patterns the models trained on different transcription levels actually capture. If the character models at the diplomatic level account for variation in spelling, it makes sense that the models would need higher order n-grams in order to capture the context of different character sequences. However, if the

models at the facsimile level capture shifts in character inventories, the context of the individual characters might be of less importance.

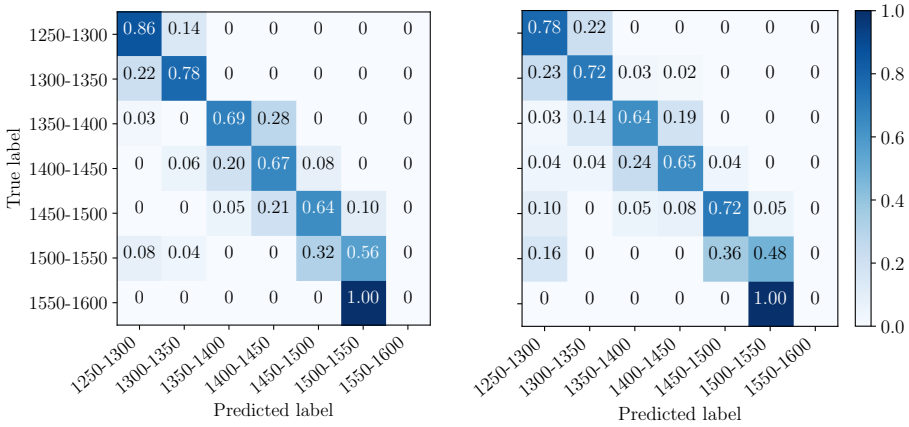
With the exception of the trigram word models, all the experiments yield higher accuracy and F1 scores than the two baselines. This suggests that the proposed models not only learn the temporal distribution of the documents over time and languages, but that they are also able to model more fine-grained temporal differences. It is also interesting to observe that the best results obtained on the 100-year bins are in line with, or for most models above, the 74.1 % state-of-the-art accuracy reported for a similar task [18]. Furthermore, although there is a decrease in accuracy when going from 100 to 50 years, our best models trained on the facsimile transcription still perform in line with or slightly better than the state-of-the-art. However, none of the models manages to predict the date of a document with a 25-year error margin with 100 % accuracy. At best, the character models using the facsimile transcription were able to correctly predict almost 75 % of the charters with a 25-year error margin. As was discussed previously, treating dating as a classification problem, in which time is viewed as a finite number of distinct classes, may yield misleading results. For example, if a charter from the very end of the 15^{th} century were assigned to the 1500-1550 time span, such a prediction would, within a classification framework, yield an accuracy of 0 % just as would be the case if a charter from the 13^{th} century were assigned to the same time span. In the former case, however, the absolute error would only be of 26 years.

Figure 4 shows confusion matrices of the cross validation errors for the highlighted models from Tables 5–8. This provides a more fine-grained view of the type of errors the models make in their predictions. The rows of the matrices represent how the documents belonging to a specific time interval were classified by the model. The numbers in the cells specify what proportion of those documents was correctly classified and what proportion was misclassified as belonging to other time spans. The individual time spans are temporally ordered along this axis. Thus, the cells on the diagonal represent the correctly classified documents within the different categories, and temporally close time spans are also closer to each other in the matrix. Firstly, a general trend across the matrices is that even when the models make a wrong prediction, in most cases, it is still a qualified guess. In fact, wrongly classified documents are mostly assigned to a time span close to the correct one. Secondly, when considering the numbers in the diagonal, it can be seen that the models have higher accuracy scores for the earlier time intervals. This is likely to be a consequence of the dataset composition, in that there are more examples from the early periods in the dataset compared to the later ones (see Table 3). Thus, while 88 % of the documents from the period 1250-1300 were correctly assigned to their time bin, this was only true for half of the documents from the period 1500-1550. Furthermore, none of documents from the period 1550-1600 was categorised correctly, but then only two charters from this period were present in the dataset used in this study.



(a) 1-3-gram char model (facsimile)

(b) 1-gram word model (facsimile)



(c) 1-3-gram char model (diplomatic)

(d) 1-gram word model (diplomatic)

Fig. 4: Confusion matrices for cross validation error in normalised counts.

7 Conclusions and future work

In this paper we investigated how state-of-the-art NLP models can be applied to the problem of dating medieval charters from 1200-to-1600 Denmark. We framed the problem as a classification task by dividing the period into bins of 50 years and using these as classes in a supervised learning setting to develop SVM classifiers. Furthermore, we investigated how different levels of transcription of the text can be used to facilitate this task.

We showed that the more detailed facsimile transcription, which captures palaeographic characteristics of a text, provided better results than the diplomatic level, where such distinctions are normalised. Moreover, both character and word n-grams showed promising results, the highest accuracy reaching 74.96 %. This level of accuracy corresponds to being able to date almost 75 % of the charters with a 25-year error margin, which philologists use as a standard for the precision with which medieval texts can be dated manually.

Looking into the accuracy results of the experiments in more depth, we showed that there was a substantial difference in how well documents from individual bins could be predicted, ranging from 88 % accuracy for the 1250-1300 documents to 52 % accuracy for documents from the years 1500-1500. We argued that this difference is likely to be due to the fact that some of the bins are represented by a few dozen documents. NLP methods often assume that a large amount of training data is available. However, small datasets are often a circumstance when working with historical text sources. In [18] the authors were able to increase the precision of their models drastically by adding synthetically generated documents to the dataset. Similar methods would be interesting to apply to our current collection. However, one danger when using such methods, is that they may be prone to overfitting. A possible way to control for this unwanted effect would be to validate the model with documents from outside the collection.

As discussed in Section 5 some charters from the medieval period are copies of earlier text making them difficult to date. In connection to the documents misclassified by the models, it would be interesting to do a more thorough inspection of these to see if some of them were indeed unidentified copies missed in the manual labelling process. The problem of outlier detection motivates another future line of work in which the temporal ranking of a collection of documents is just as important as the actual dating.

In this paper we compared the different models by looking at their performance measured in terms of accuracy. We haven't yet, however, investigated the behaviour or the individual predictions made by the different models. The interpretability of machine learning models is currently a much debated topic among NLP researchers, the motivation for this line of work being the importance of creating trust in the models' predictions and a wish to infer causality in the natural world from synthetic learning settings [8].

Both causality and trust are relevant when studying automatic methods for the dating of historical documents: Considering causality first, the ability to answer questions about what types of feature were relevant when using the fac-

simile and diplomatic transcriptions, respectively, would help us answer questions about how the charters developed over time. Similarly, when comparing word and character models, it would be useful to determine if the most predictive features reflect lexical, phonological and morphological characteristics of the manuscripts that a philologist would recognise as being relevant to their dating. As for trust, it is relevant when applying trained models to undated documents. Knowing why the models fail or succeed is a crucial step if we wish to apply these models to currently undated documents, such as the copies we mentioned earlier. Being able to interpret such models might in turn contribute to studies of the diachronic developments within text collections using different linguistic annotations as basis for such analyses.

In this study we looked at two different levels of transcription of the charters, one that captured palaeographic characteristics and the other where such differences were normalised. In the future it would be interesting to repeat the study with other levels of transcriptions, e.g., the normalised level, or to include different types of linguistic annotation such as POS tags or morphological features. Moreover, as mentioned earlier, it is not only the text that provides clues to when a manuscript was written, but so does physical evidence about ink and parchment. In this respect a line of future work could be to investigate how methods of ensemble learning can contribute to the problem of dating documents, by combing the textual evidence outlined in this paper with evidence from image processing or multispectral measurements of the material.

References

1. Abe, H., Tsumoto, S.: Text categorization with considering temporal patterns of term usages. In: Proceedings of the 2010 IEEE International Conference on Data Mining Workshops. pp. 800–807. ICDMW '10, IEEE Computer Society, Washington, DC, USA (2010). https://doi.org/10.1109/ICDMW.2010.186
2. Garcia-Fernandez, A., Ligozat, A.L., Dinarelli, M., Bernhard, D.: When was it written? Automatically determining publication dates. In: Grossi, R., Sebastiani, F., Silvestri, F. (eds.) String Processing and Information Retrieval. pp. 221–236. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
3. Hansen, A.: Adkomstbreve i Skt. Clara Klosters arkiv. In: Driscoll, M., Óskarsdóttir, S. (eds.) 66 håndskrifter fra Arne Magnussons samling, pp. 138–139. Museum Tusculanum (2015)
4. Haugen, O., Bruvik, T., Driscoll, M., Johansson, K., Kyrkjebø, R., Wills, T.: The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources. The Medieval Nordic Text Archive, 2 edn. (2008)
5. He, S., Sammara, P., Burgers, J., Schomaker, L.: Towards style-based dating of historical documents. In: 14th International Conference on Frontiers in Handwriting Recognition. pp. 265–270 (Sep 2014). https://doi.org/10.1109/ICFHR.2014.52
6. He, S., Schomaker, L.: A polar stroke descriptor for classification of historical documents. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 6–10 (Aug 2015). https://doi.org/10.1109/ICDAR.2015.7333715

7. Kjeldsen, A.S.: Filologiske studier i kongesagahåndskriftet Morkinskinna, Bibliotheca Arnamagnæana Supplementum, vol. 8. Museum Tusculanums Forlag, København (2013)
8. Lipton, Z.C.: The mythos of model interpretability. arXiv:1602.04938v3 [cs.LG] (2016)
9. Medvedeva, M., Kroon, M., Plank, B.: When sparse traditional models outperform dense neural networks: The curious case of discriminating between similar languages. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). pp. 156–163 (2017)
10. Pichel Campos, J.R., Gamallo, P., Alegria, I.: Measuring language distance among historical varieties using perplexity. Application to European Portuguese. In: Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018). pp. 145–155. Association for Computational Linguistics (2018), http://aclweb.org/anthology/W18-3916
11. Popescu, O., Strapparava, C.: Behind the times: Detecting epoch changes using large corpora. In: International Joint Conference on Natural Language Processing IJCNLP. pp. 347–355. Nagoya, Japan (October 14-18 2013)
12. Popescu, O., Strapparava, C.: Semeval 2015, task 7: Diachronic text evaluation. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 870–878 (2015)
13. Štajner, S., Zampieri, M.: Stylistic changes for temporal text classification. In: Habernal, I., Matoušek, V. (eds.) Text, Speech, and Dialogue. pp. 519–526. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
14. Szymanski, T., Lynch, G.: Ucd: Diachronic text classification with character, word, and syntactic n-grams. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 879–883 (2015)
15. Wahlberg, F., Mårtensson, L., Brun, A.: Large scale continuous dating of medieval scribes using a combined image and language model. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS). pp. 48–53 (Apr 2016). https://doi.org/10.1109/DAS.2016.71
16. Wahlberg, F., Mårtensson, L., Brun, A.: Large scale style based dating of medieval manuscripts. In: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing. pp. 107–114. HIP '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2809544.2809560
17. Wahlberg, F., Wilkinson, T., Brun, A.: Historical manuscript production date estimation using deep convolutional neural networks. In: 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 205–210 (Oct 2016). https://doi.org/10.1109/ICFHR.2016.0048
18. Zampieri, M., Malmasi, S., Dras, M.: Modeling language change in historical corpora: The case of Portuguese. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). pp. 4098–4104. European Language Resources Association (ELRA), Paris, France (May 2016)
19. Zampieri, M., Malmasi, S., Nakov, P., Ali, A., Shon, S., Glass, J., Scherrer, Y., Samardžić, T., Ljubešić, N., Tiedemann, J., van der Lee, C., Grondelaers, S., Oostdijk, N., van den Bosch, A., Kumar, R., Lahiri, B., Jain, M.: Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In: Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). Santa Fe, USA (2018)