Survey and reproduction of computational approaches to dating of historical texts

Sidsel Boldsen
Dept. of Nordic Studies and Linguistics
University of Copenhagen
Denmark
sbol@hum.ku.dk

Fredrik Wahlberg
Dept. of Linguistics and Philology
Uppsala University
Sweden
fredrik.wahlberg@lingfil.uu.se

Abstract

Finding the year of writing for a historical text is of crucial importance to historical and philological research. However, the year of original creation is rarely explicitly stated and must be inferred from the text content, historical records, and codicological clues. Given a transcribed text, machine learning has successfully been used to estimate years of production. In this paper, we present an overview of estimation approaches from the literature for historical text archives, spanning from the 12^{th} century until today.

1 Introduction

Knowing when a text was written is of crucial importance for relating its content to a historical context. With the increasing digitization of historical archives, many new research opportunities have emerged for studying how languages have evolved. However, such studies rely on digitized corpora explicitly stating when the texts were originally written. This information is often not given by the original scribe, although educated guesses from later owners can sometimes be found in manuscripts. Additionally, improved dating of historical manuscripts can help historians to better understand the chronology of their sources.

The premise for our paper is an imagined scenario where a historian or philologist needs help with a transcribed collection. We imagine being given a partially annotated set of documents (given either as specific years or as intervals) and employing a computer model to determine the production years of the unlabelled documents. Although there is literature describing different ways of solving the problem of the above scenario, there is little work done on comparing the different modelling approaches. In this paper, we will survey and evaluate computational approaches to the problem of estimating the production dates of text in digitalized historical archives. We have reimplemented several methods for estimation and feature extraction proposed in the literature. Our experimental setup allows us to evaluate combinations of different methods on datasets representing different times, text lengths, and genres. Our reimplementation is available as open source[1].

Our primary historical datasets were two medieval archives containing legal documents from Denmark and Sweden. Comparing results on these collections is of special interest, as they are similar with respect to content, but differ in the number of documents, temporal distributions and detail of annotation. To assess the generalizability of the methods we also include two modern collections. These modern collections, that have previously been treated in the literature, are a collection of English news items, from the SemEval 2015 shared task on diachronic text evaluation (Popescu and Strapparava, 2015), and Colonia, a corpus of historical Portuguese (Zampieri and Becker, 2013).

An overview of the relevant literature is presented in Section 2, Section 3 contains a description of the datasets, our experimental setup is described in Section 4, and, finally, results and discussion are presented in Section 5.

[1] IPython notebooks can be found at http://github.com/fredrikwahlberg/nodalida21

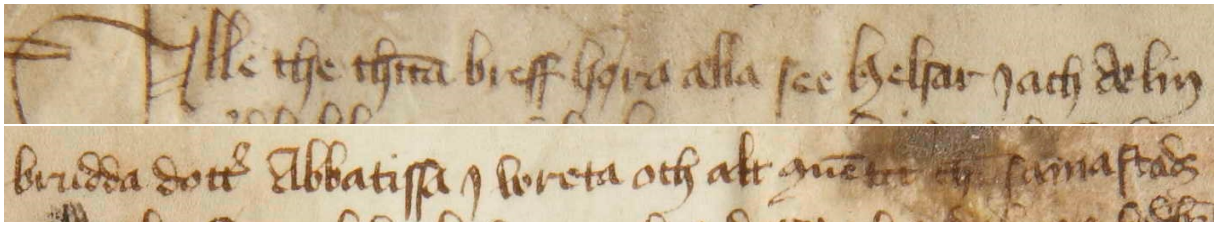

Figure 1: The first line of charter SDHK 18863, containing an agreement on an exchange of land in 1417. The text is "Alle the thetta breff høra ælla see helsar jach Ælin Bruddadotter, abbatissa i Wreta, och alt conuentit ther samastadz" (from "Svenskt Diplomatariums Huvudkartotek" 18863, section 3).

2 Previous work

The problem of automatic text dating has been treated using various methods and applied to a wide range of different types of corpora. To the best of our knowledge, the task of assigning a date to documents was first introduced in the information retrieval community with the main goal to query document collections based on temporal relevance. De Jong et al. (2005) treat the problem as a text classification task in which documents are dated by comparing them to temporal query profiles. They refer to such profiles as *temporal language models*, which essentially capture the distribution of term or concept usage over time. The same idea is found in the work of Dalli and Wilks (2006), in which word frequencies across time are used to infer temporal association rules. The work by de Jong et al. (2005) was later expanded by Kanhabua and Nørvåg (2008) who improved the temporal language models by applying various steps of pre-processing, including filtering words based on TF-IDF scores and POS tags, applying stemming, and collocation extraction.

The above works were made on corpora of newspaper articles, which have remained to be an object of research within the field of automatic dating, most recently in the SemEval 2015 shared task on diachronic text evaluation (DTE) on English news snippets (Popescu and Strapparava, 2015). Following the work on temporal language modelling, Garcia-Fernandez et al. (2011) introduced using support vector machines (SVM) for the task of dating, in which documents are represented by feature vectors of word and character counts, in addition to other handcrafted features. Whereas in the temporal language modelling approach the sole problem is to learn the distribution of words in a set of documents belonging to a specific time span, the goal of mapping a document to a date is now part of the learning objective. Later work on news corpora has focused on how the extraction of temporal references, such as expressions for time and events, can facilitate the task of dating, which was also the research question in two of the three subtasks of the DTE shared task (Chambers, 2012; Vashishth et al., 2019).

Aside from news, scholars have studied a wide range of different historical corpora, ranging from broad collections such as Google n-grams (Popescu and Strapparava, 2014) to more narrow collections as in the DaDoE-val2020 shared task (Menini et al., 2020), which introduced a diachronic corpus of political work by Alcide De Gasperi. While news items naturally contain explicit temporal references for when the text was written, this is often not the case when working with other genres. For example, if a philologist were to date a piece of literature, their work may solely rely on features such as lexicon, grammar, topic, or style, as the contemporary context is often implicit. Thus, work outside the news genre has generally put less emphasis on extracting temporal references, and instead explore how the language in a text can be represented.

One of the first studies to extend the work beyond the news genre was Kumar et al. (2011), who use language modelling to predict the date of a collection of short stories published between 1798 to 2008 from Project Gutenberg[2]. Subsequently, language modelling has not been applied to the problem of

[2] https://www.gutenberg.org

dating. Work has been done to identify temporal trends in historical corpora (Pichel Campos et al., 2018; Pichel et al., 2020; Boldsen et al., 2019), by using language modelling to measure the distance between time periods, but models were not explicitly applied to the task of dating. Instead, studies have focused on creating vector representations of documents and then using those for classification. The raw text has been used directly as input to create bag-of-words and/or characters, with n-gram sizes ranging from one to three words (Niculae et al., 2014; Szymanski and Lynch, 2015; Zampieri et al., 2016), and one to five characters (Garcia-Fernandez et al., 2011; Niculae et al., 2014; Szymanski and Lynch, 2015). Other features may be extracted, such as syntactic features using POS annotations (Szymanski and Lynch, 2015; Zampieri et al., 2015, 2016) and stylistic measures such as lexical diversity (Štajner and Zampieri, 2013; Zampieri et al., 2015).

Most commonly, the problem of dating a text is defined as a classification problem in which classes are treated as bins corresponding to different time spans. Several estimators have been applied, including logistic regression (Chambers, 2012), support vector machines (Garcia-Fernandez et al., 2011; Szymanski and Lynch, 2015; Zampieri et al., 2016) and multinomial naive Bayes (Mihalcea and Nastase, 2012; Zampieri et al., 2016). The size of the bins depend on the problem and the data available. For dating of contemporary news items, scholars have worked with granularities down to a yearly basis (Chambers, 2012; Vashishth et al., 2019). This is typically not possible when working with historical text, as data is sparse and may in turn not have such a precise date of production. Here, scholars have instead worked on dating documents within a century (Štajner and Zampieri, 2013) or a decade (Popescu and Strapparava, 2015).

Compared to classification, regression methods have not been extensively explored. In regression, samples are mapped to a date directly instead of a bin, thus circumventing the obstacle of deciding on a specific bin size. Also, regression preserves the ordinal nature of the problem, which classification ignores. Niculae et al. (2014) propose to use ordinal regression. In this approach, the task of dating is considered as a ranking problem, where reference documents are placed on a timeline, which is then used to estimate the most probable time spans for query documents. Another attempt using regression comes from the field of image processing, where Wahlberg et al. (2016) applied Gaussian Processes (GP) to the problem of estimating the date of medieval manuscripts using visual features extracted from the facsimile together with the transcribed text.

Whether classification or regression is best suited for the problem of dating text - and what pitfalls such approaches have - are still open questions. As for feature extraction, neural methods have over the last decades undermined the use of manual feature extraction for a wide range of problems, including text classification. Vashishth et al. (2019) applied graph convolutional networks to the problem of dating, utilizing syntactic information and temporal reference extraction in addition to the words of the text. For smaller corpora, neural approaches are yet to be tested, which is out of the scope of this paper. Instead, we seek to describe and compare the methods that have already been established for the dating of historical text corpora.

3 Datasets

3.1 Svenskt Diplomatariums Huvudkartotek

"Svenskt Diplomatariums Huvudkartotek" (SDHK) is a collection of charters from medieval Sweden (c. 1050-1523). The collection consists of approximately 44,000 charters on (mostly) parchment, of which about 10,500 have been transcribed. The most frequent languages are Swedish (c. 3,000 transcribed charters) and Latin (c. 7,500 transcribed charters). Of the full collection, about 11,000 charters have been photographed, largely overlapping with the transcribed set.

While the Latin vocabulary and spelling are fairly consistent, except for small variations in the use of abbreviations, the Swedish text changes significantly with time. The Swedish language goes through significant development from Old Swedish ("fornsvenska") involving grammar, lexicon, and spelling between the 13^{th} and 16^{th} centuries. The material is fur-

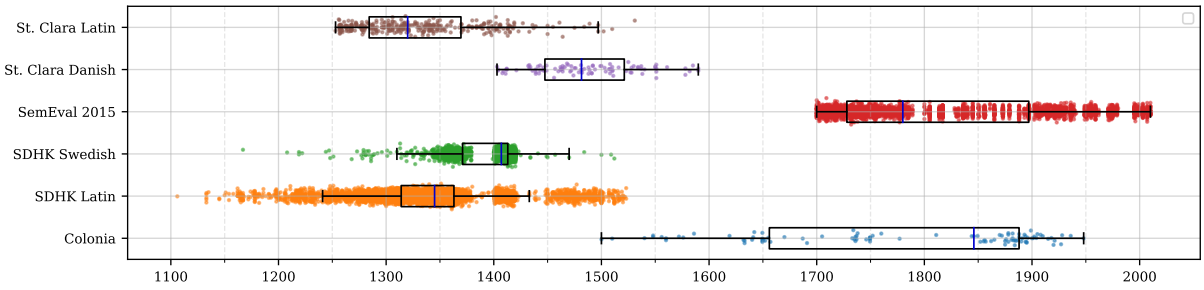

Figure 2: Box plots showing the represented years for the document collections (coloured dots are individual documents) in each dataset described in section 3.

ther complicated by the transcribers' inconsistent expansion of abbreviations and spelling normalization. Since these types of problems are common with this type of archive, we did not see any need for further annotation or human curation. Hypothetically, if many researchers, each primarily interested in their limited period, have contributed to the transcribed collection, then transcription standards (e.g., expansions of abbreviations) might change over time, through not due to changes in the underlying historical material. Though a machine could potentially overfit on such features, we see this as falling outside the scope of this paper.

3.2 The charters of St. Clara Convent

The charters of St. Clara Convent (Roskilde, Denmark) are part of the Arnamagnæan Collection at the University of Copenhagen (Hansen, 2015). The charters date from when the convent was founded in 1256 till it was closed after the Reformation, after which the convent and its archive became part of the university's properties. In total, 471 charters are left from the old archive. The majority of the charters are written in Latin and Danish (361 and 100 charters, respectively), the rest being in German or Swedish.

The charters have been all digitized with multiple layers of annotation, including both a *facsimile* and *diplomatic* transcription of the text. The facsimile level (a) captures the handwritten form of the text by annotating the palaeographic characteristics of the letters (i.e., focusing on the shape of the character rather than solely on its meaning). The diplomatic transcription (b) is of the kind that is usually found in manuscript editions. At this level, the difference in handwriting is ignored and abbreviated diacritics are expanded, while variation in spelling is still preserved:

(a) ſoꝛoꝛ ⁊ ꝯonaſteꝛíj earu̇ ın poſterum

(b) soror(um) (et) monasterij earu(m) in posterum

In the example above, the word "ſoꝛoꝛ" is written in a way very similar to the original handwriting (i.e., as a facsimile). In diplomatic annotation, this becomes "soror(um)", where "soror" are the modern forms of the letters and the "-um" suffix is expanded from the stroke on the last letter and inferred from the context.

3.3 SemEval2015

The SemEval2015 shared task of "Diachronic Text Evaluation" introduces a corpus of English news snippets dating from the 18th to the 21th centuries (Popescu and Strapparava, 2015). Contrary to the collections of charters, the news snippets were not precisely dated but rather given as intervals over years (2 and 6 years wide) which can be seen by the distribution of data points in Figure 2. For this paper, we only utilize the training set data from the task, ending up with c. 4,500 documents.

3.4 Colonia

Colonia is a corpus of historical Portuguese, compiled from various sources spanning from the 16th to the 20th century (Zampieri and Becker, 2013). While the collection of news snippets and charters contains text with lengths ranging from 10 to hundreds of words, the texts of Colonia are substantially longer, containing full works with thousands of tokens.

Thus, with the 100 documents that it contains, the collection counts up to five million tokens in total.

4 Experimental setup

In our experimental setup, we have implemented a number of ways of doing feature extraction. We then evaluated all combinations between those feature spaces and a number of ways of doing the mapping to years on the timeline.

In the documents of several datasets, clearly stated years can be found. In order not to let the estimators simply learn to find this information (especially in the charter datasets, where Roman numerals are frequently encountered) we have removed all numerals from the text as a part of the preprocessing.

Some of the methods we have evaluated were quite demanding of the hardware. Hence, we randomised the training, validation, and test sets while preprocessing the datasets, using the same sets for all evaluations. It should be noted that this is not standard for several of our approaches (e.g., naive Bayes) which are normally evaluated using cross-validation. However, we saw this as the only way of making a fair comparison and not risk giving some estimators more or different data.

4.1 Feature spaces

Binary bag-of-words vectors (BOW) (i.e., encoding the existence or absence of a word) have been shown to be useful in many applications. Since the popularity of words changes over time, this type of vector can encode distributional information on word choice. We generated such vectors from the training and validation folds of the datasets and then transformed the full datasets into their respective vector space representations. This meant that only the part of the test set vocabulary that was overlapping with the training and validation vocabularies was used. As the Colonia dataset had a higher level of annotation, we made binary BOW vectors from the words, pos-tags, and concatenated word+pos-tags.

Several papers use n-gram feature vectors on both the word and character level. Looking at a small context around words has the potential to encode changes in common expressions or even some semantics. In contrast, character level n-grams have the potential to catch spelling or phonetic changes (especially during eras where there were no standardised spellings). The order of a space spanned by n-grams is only limited by computer memory. We chose to extract n-grams of orders $\{1, 2, 3\}$.

For some estimators, it is considered best practice to perform feature selection as noise removal and to lower run times. We ran feature selection based on χ^2 statistics, capping the feature space dimensionality to 1000 for all estimators, but only kept the automatic selection for those estimators where the training accuracy improved (Gaussian process, linear SVM, and non-linear SVM).

4.2 Classification for date estimation

Usually, the date estimation was treated as a classification task in the literature. This was done by formulating the mapping from documents to the timeline by dividing the timeline 25-year wide bins and then classifying the documents into those bins. An advantage of this approach was that several estimators can be used, specializing on particular parts of the timeline (Garcia-Fernandez et al., 2011; Zampieri et al., 2016).

The most popular estimation method in our chosen literature is the support vector machine (SVM) (Cortes and Vapnik, 1995). One core advantage with the SVM is that finding a separation in some feature space is a reasonable fast convex optimisation problem. The resulting linear decision boundary is interpretable in term of the feature set, especially with BOW vectors, but suffers from the fact that the data needs to be linearly separable. In the literature, strategies for finding hyper-parameters or kernels are surprisingly absent. From this, we draw the conclusion that (most likely) a linear SVM was used, which only has one regularisation hyper-parameter. Because of the high dimensionality of some feature spaces, a non-linear decision boundary is often not needed (and expensive). For testing this in our setting, we extended our experiments by using the standard radial basis function (RBF) kernel to introduce some non-linearity, in addition to the linear SVM.

Temporal language models are probabilistic models over sequences of tokens, either words

or characters, for a given set of time spans. The model approximates the likelihood of a sequence, given some corpus. To simplify such models, the Markov assumption is commonly used to split up longer sequences, creating a so-called n-gram model (as in the feature described above). In order to create temporal language models for classification, we split up the data into bins and trained language models on these respective bins (Boldsen and Paggio, 2019). Given this set of temporal language models, dating a document is equivalent to finding the model that is more likely to generate a specific document. One of the issues in estimating sequence probabilities is encountering unseen n-grams. This is commonly handled by modifying the n-gram counts by discounting from non-zero events. In this paper, we used modified Kneser-Ney smoothing with interpolation (Chen and Goodman, 1999).

Naive Bayes classifiers (surprisingly) often deliver good results in a variety of domains despite their assumption of independence between features. Zampieri et al. (2016) employ a multinomial naive Bayes classifier, which is common for linguistic applications. This fits well with their chosen feature model, focusing on the frequencies of words and POS-tags. For the completeness of the comparison, we evaluate estimators using both multinomial and Gaussian priors.

4.3 Regression for date estimation

To get around the problem of choosing the proper bin width for a classification, some papers treat dating as a regression problem. In Wahlberg et al. (2016), a Gaussian process (GP) was used for the regression, allowing mapping from documents to normal distributions over the timeline (i.e., inferring uncertainties in addition to point estimates).

For a GP, the weight vector ω, in the standard regression expression $\hat{y}_i = \omega\phi(x_i)$, is treated as a random vector from a multivariate normal distribution (Rasmussen and Williams, 2006). Though the GP is non-parametric and ω is analytically inferred from the data, the hyper-parameters for the feature transform (kernel) $\phi(\cdot)$ must be trained (we used RBF as to be able to compare to the SVM) by maximizing the likelihood of generating the training data given that parameter set. Since GPs are generative and probabilistic, all hyperparameters can be marginalized. However, this is rarely done in practice. Most often, a set of hyperparameters are chosen by maximizing their likelihood given the model (maximum a posteriori).

	MVB	Uniform	Weighted
Colonia	26.32	5.65	11.74
SDHK Latin	26.29	5.89	17.69
SDHK Swedish	69.04	7.16	54.24
SemEval 2015	23.64	7.72	11.32
St.Clara dipl. Danish	15.79	12.48	14.27
St.Clara dipl. Latin	19.72	8.31	14.71
St.Clara facs. Danish	15.79	12.49	13.85
St.Clara facs. Latin	19.72	8.25	14.68

Table 1: Accuracy for different baseline strategies. The majority vote baseline (MVB) classifies all documents as the most common class while the other baselines are expected accuracy with hypothetical random classifier. The "uniform" baseline classifier draws random years from a uniform distribution over the relevant timeline, while the "weighted" draws from each dataset's label distribution.

4.4 Evaluation metric

In most of the papers presented in Section 2, accuracy was the preferred evaluation metric. For any classification over a timeline, a bin width needs to be chosen. Several papers used 50-year-wide non-overlapping bins. In our implementation we have chosen 25-year wide bins, making accuracy less forgiving.

As for accuracy baselines, we created random baseline classifications using three strategies. First, the majority vote baseline (i.e., always classifying as the most common bin), a uniform bin probability, and a weighted scheme with random classifications while respecting the date distribution of the data. The baseline accuracy scores can be found in Table 1. Given these methods, any accuracy above 25% can be seen as better than random for all datasets except for SDHK in Swedish, which is heavily skewed and has a majority vote baseline of 69%.

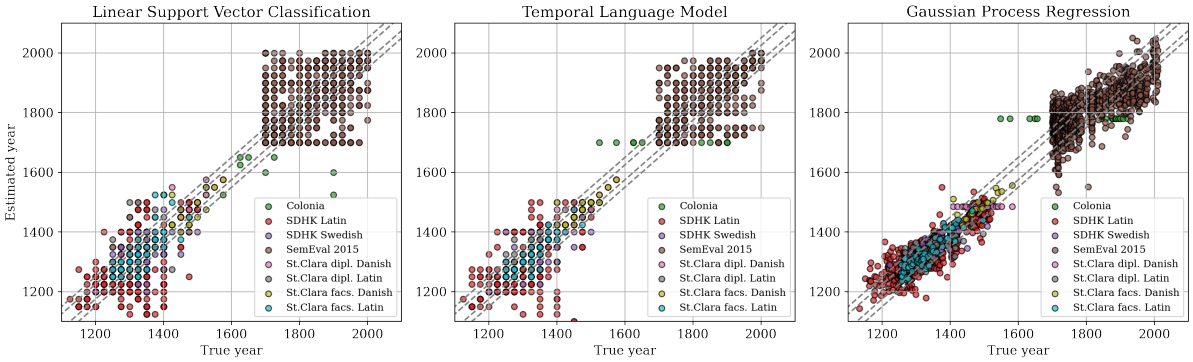

Figure 3: Scatter plot over the estimated production years versus their true years for three types of estimators. All used character bigram features and classification except for the rightmost that used regression. The dashed lines on the diagonal are spaced 25 and 50 years from the diagonal. Note that many points are plotted on top of each other, especially for the classification based estimators.

5 Results and discussion

The results from the experiments can be found in Table 2. Due to a lack of space, we chose to focus the discussion on the four different classifiers that provide the highest scores for the individual datasets (highlighted with red). The results for the remaining two classifiers can be found in the appendix.

All classifiers perform above baseline for at least one feature set. Considering the best performing feature sets per classifier (highlighted with blue), character models perform the best across classifiers except for Gaussian Naive Bayes. Aside from being able to capture features such as morphology and spelling, character models have the advantage that the feature space is smaller than for word models, which in turn increases the number of examples that estimators consider. Whether it is the features or simply the data size that is at play is difficult to read from these numbers.

When working with vector representation of words and higher level character n-grams, the feature set easily becomes larger than the number of samples used for training a model. In these cases, one could argue that it is unlikely for the estimators using these representations (SVMs, naive Bayes) *not* to find something in the training set that correlates with the timeline, even though the feature might not necessarily be related to language change. The problem is compounded by that the training, validation, and test data were all drawn from the same data generating process and, hence, might have the same spurious correlations in relation to the target labels.

If we compare the linear SVM with the non-linear SVM, the linear version has the advantage of being more qualitatively interpretable due to the lack of warping of the feature space. However, if we compare the models in terms of accuracy, using a non-linear kernel yields slightly better results. When we compare the test set predictions of the different estimators, they do tend to correlate. As is revealed in Figure 4, there is a strong relationship between the predictions made using different SVM estimators (linear and non-linear), especially on similar feature sets. If we consider the predictions using the non-linear SVM on character unigrams, we see a slightly stronger correlation with the predictions of the linear SVM when using higher orders, which suggests that a more complex model is able to utilize its non-linear combination of features on the problem. However, in terms of accuracy results, this advantage is not widely outspoken. Thus, we argue that choosing a linear kernel may still be preferable, as its predictions are more easily explained to a community of philologists or historians.

Despite not appearing in more recent research, the temporal language model outperforms other models on several datasets using character features. All estimators that we have evaluated describe language as a distribution of words or characters. What distin-

	char$_1$	char$_2$	char$_3$	word$_1$	word$_2$	word$_3$
	Temporal Language Model Classification					
Colonia	5.3	5.3	5.3	5.3	5.3	5.3
SDHK Latin	0.5	69.2	75.3	0.0	7.5	15.7
SDHK Swedish	1.9	93.5	95.0	0.3	1.0	5.7
SemEval 2015	25.7	45.7	58.3	1.2	15.6	16.5
St. Clara dipl. Danish	15.8	42.1	31.6	0.0	31.6	31.6
St. Clara dipl. Latin	1.4	54.9	56.3	0.0	26.8	29.6
St. Clara facs. Danish	42.1	63.2	57.9	5.3	10.5	10.5
St. Clara facs. Latin	7.0	69.0	71.8	0.0	1.4	2.8
	Linear Support Vector Classification					
Colonia	36.8	47.4	36.8	36.8	42.1	36.8
SDHK Latin	37.5	53.9	53.4	41.3	35.1	34.2
SDHK Swedish	81.0	89.0	88.0	76.2	69.9	69.4
SemEval 2015	26.8	31.4	30.2	28.4	24.7	24.2
St. Clara dipl. Danish	26.3	57.9	36.8	10.5	10.5	10.5
St. Clara dipl. Latin	38.0	47.9	39.4	32.4	19.7	26.8
St. Clara facs. Danish	42.1	31.6	10.5	0.0	10.5	21.1
St. Clara facs. Latin	47.9	49.3	36.6	33.8	15.5	22.5
	Gaussian naive Bayes					
Colonia	31.6	21.1	31.6	26.3	36.8	42.1
SDHK Latin	12.9	37.2	58.3	62.9	-	-
SDHK Swedish	19.8	84.6	92.7	86.9	88.8	-
SemEval 2015	19.7	21.7	39.8	50.9	49.0	43.1
St. Clara dipl. Danish	31.6	26.3	21.1	36.8	47.4	15.8
St. Clara dipl. Latin	16.9	39.4	54.9	54.9	66.2	69.0
St. Clara facs. Danish	26.3	31.6	36.8	36.8	47.4	36.8
St. Clara facs. Latin	53.5	67.6	70.4	63.4	66.2	59.2
	Support Vector Classification with Radial Basis Function					
Colonia	42.1	52.6	42.1	42.1	42.1	42.1
SDHK Latin	45.0	53.4	58.3	48.0	40.1	10.6
SDHK Swedish	88.3	90.3	90.1	80.6	1.3	1.5
SemEval 2015	27.6	30.1	31.6	27.4	10.3	14.3
St. Clara dipl. Danish	26.3	57.9	26.3	21.1	15.8	21.1
St. Clara dipl. Latin	45.1	46.5	50.7	38.0	25.4	19.7
St. Clara facs. Danish	36.8	21.1	31.6	21.1	15.8	26.3
St. Clara facs. Latin	50.7	47.9	45.1	33.8	15.5	25.4

Table 2: The accuracy scores (in percent) for the four estimators. Best results for each dataset are highlighted with red, and best results for each estimator are highlighted with blue. We ran several more combinations of feature sets and estimators, all of which can be found in our code repository for this paper.

guishes the temporal language modelling approach from the other estimators, is that it uses perplexity as a measure to model linguistic difference. Several estimators are treating the probability density functions for the different documents as points in a Euclidean space (e.g., linear SVM). This assumption often works. However, by using a divergence metric between probability density functions, the space is treated more in line with the nature of the encoding. This has been shown to be beneficial for image based dating of manuscripts (Wahlberg et al., 2014), leading us to speculate that this result is valid here too.

While performing well on character feature sets, the temporal language model struggles when it comes to word representations with accuracies below 10%. This suggests that the temporal language model is sensitive to larger feature spaces, in which smoothing might not be sufficient. Furthermore, it performs poorly on the Colonia dataset. Whether this is due to the number of samples, document length, or dataset distribution is difficult to say, and it calls for further analysis of the models with respect to dataset statistics.

Finally, we wish to discuss the performance of regression to classification methods. Most previous work has preferred to use classifica-

Figure 4: A heat map of the correlation coefficients ($p < 0.005$) between test set predictions by SVM estimators with linear and RBF kernels using different feature sets. The coefficients ($p < 0.005$) were computed as Kendall's τ, which does not assume a normal distribution and works for ordinal values.

tion instead of regression, treating the timeline as discrete and with temporally independent labels. That labels are independent is reflected in the use of categorical accuracy as the evaluation metric. If we look at Figure 3, this is illustrated by the inner dashed lines, outside which predictions are considered incorrect, even though they are close to the target temporally. In this respect, regression methods should have an advantage, however, this is not reflected in our results. It would be interesting to further compare what advantages there are - if any - to choosing regression over the classification.

6 Conclusion and Future Work

In this paper, we present a survey of several methods found in the literature for estimating the production years of transcribed historical documents. We have reproduced the methods used in a number of papers, including different n-gram/word/pos-tag feature spaces and several linear (naive Bayes, linear SVM) and non-linear (Gaussian process, SVM with RBF kernel) estimators.

Our results show that several of the combinations of estimators and feature models work well, but that character n-gram features provide the best results overall. In particular, the temporal language model with character features surpasses more recently proposed models. Whether this is due to the linguistic features (e.g., suffixes or phonetic changes leading to changes in spelling) that they potentially capture or simply due to a reduced feature space giving better model parameter estimates, we cannot conclude from our results. Therefore, we call for further analysis of the estimators, preferably favouring more interpretable approaches (e.g., linear SVM).

Our experiments show that combinations of estimators and feature transforms that worked well on younger materials were often also successful on older materials, and vice versa. As the datasets that we compare not only differ in age, but also in number and size of samples. For future work, it would be interesting to investigate the robustness of the methods from the literature with respect to such dataset statistics. In this respect, it would also be relevant to include recent work on neural models such as using word embeddings and convolutional networks, which have been shown to work well for dating on large corpora. However, these have yet to be trialed on smaller corpora.

Acknowledgments

The first author is supported by the project *Script and Text in Time and Space*, a core group project supported by the Velux Foundations. We are grateful to Patrizia Paggio for her support and comments regarding this paper.

We also want to thank the Swedish National Archive for providing the SDHK dataset, both as images and transcribed text. Funding for the second author was provided by the project "New Eyes on Sweden's Medieval Scribes", headed by Lasse Mårtensson.

Finally, we want to thank the anonymous reviewers for finding the time to give constructive criticism.

References

Sidsel Boldsen, Manex Agirrezabal, and Patrizia Paggio. 2019. Identifying temporal trends based on perplexity and clustering: Are we looking at language change?

Sidsel Boldsen and Patrizia Paggio. 2019. Automatic dating of medieval charters from denmark. In *DHN*.

Nathanael Chambers. 2012. Labeling documents with timestamps: Learning from their time expressions. Technical report, NAVAL ACADEMY ANNAPOLIS MD DEPT OF COMPUTER SCIENCE.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394.

Corinna Cortes and Vladimir Vapnik. 1995. https://doi.org/10.1023/A:1022627411411 Support-vector networks. *Machine Learning*, 20(3):273–297.

Angelo Dalli and Yorick Wilks. 2006. Automatic dating of documents and temporal text classification. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 17–22.

Anne Garcia-Fernandez, Anne-Laure Ligozat, Marco Dinarelli, and Delphine Bernhard. 2011. When was it written? Automatically determining publication dates. In *String Processing and Information Retrieval*, pages 221–236, Berlin, Heidelberg. Springer Berlin Heidelberg.

Anne Mette Hansen. 2015. Adkomstbreve i Skt. Clara Klosters arkiv. In Matthew J. Driscoll and Svanhildur Óskarsdóttir, editors, *66 håndskrifter fra Arne Magnussons samling*, pages 138–139. Museum Tusculanum.

Franciska de Jong, Henning Rode, and Djoerd Hiemstra. 2005. Temporal language models for the disclosure of historical text. In *Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*, pages 161–168. Koninklijke Nederlandse Academie van Wetenschappen.

Nattiya Kanhabua and Kjetil Nørvåg. 2008. Improving temporal language models for determining time of non-timestamped documents. In *International Conference on Theory and Practice of Digital Libraries*, pages 358–370. Springer.

Abhimanu Kumar, Matthew Lease, and Jason Baldridge. 2011. Supervised language modeling for temporal resolution of texts. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2069–2072.

Stefano Menini, Giovanni Moretti, and S. Tonelli R. Sprugnoli. 2020. Dating document evaluation at EVALITA 2020. `https://dhfbk.github.io/DaDoEval/`. Accessed: 2020-08-03.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 259–263.

Vlad Niculae, Marcos Zampieri, Liviu P Dinu, and Alina Maria Ciobanu. 2014. Temporal text ranking and automatic dating of texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 17–21.

José Ramom Pichel, Pablo Gamallo, Iñaki Alegria, and Marco Neves. 2020. https://doi.org/10.1080/09296174.2020.1732177 A methodology to measure the diachronic language distance between three languages based on perplexity. *Journal of Quantitative Linguistics*, 0(0):1–31.

José Ramom Pichel Campos, Pablo Gamallo, and Iñaki Alegria. 2018. http://aclweb.org/anthology/W18-3916 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)*, pages 145–155. Association for Computational Linguistics.

Octavian Popescu and Carlo Strapparava. 2014. Time corpora: Epochs, opinions and changes. *Knowledge-Based Systems*, 69:3–13.

Octavian Popescu and Carlo Strapparava. 2015. Semeval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 870–878.

C. E. Rasmussen and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Sanja Štajner and Marcos Zampieri. 2013. Stylistic changes for temporal text classification. In *Text, Speech, and Dialogue*, pages 519–526, Berlin, Heidelberg. Springer Berlin Heidelberg.

Terrence Szymanski and Gerard Lynch. 2015. UCD: Diachronic text classification with character, word, and syntactic n-grams. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 879–883.

Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha Talukdar. 2019. http://arxiv.org/abs/1902.00175 Dating documents using graph convolution networks.

F. Wahlberg, L. Mårtensson, and A. Brun. 2014. https://doi.org/10.1109/ICFHR.2014.128 Scribal attribution using a novel 3-d quill-curvature feature histogram. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pages 732–737.

F. Wahlberg, L. Mårtensson, and A. Brun. 2016. Large scale continuous dating of medieval scribes using a combined image and language model. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 48–53.

Marcos Zampieri and Martin Becker. 2013. Colonia: Corpus of historical portuguese. *ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research*, 5:69–76.

Marcos Zampieri, Alina Maria Ciobanu, Vlad Niculae, and Liviu P Dinu. 2015. Ambra: A ranking approach to temporal text classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 851–855.

Marcos Zampieri, Shervin Malmasi, and Mark Dras. 2016. Modeling language change in historical corpora: The case of Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4098–4104, Paris, France. European Language Resources Association (ELRA).

A Experimental results

	Temporal Language Model Classification					
	char$_1$	char$_2$	char$_3$	word$_1$	word$_2$	word$_3$
Colonia	5.3	5.3	5.3	5.3	5.3	5.3
SDHK Latin	0.5	69.2	75.3	0.0	7.5	15.7
SDHK Swedish	1.9	93.5	95.0	0.3	1.0	5.7
SemEval 2015	25.7	45.7	58.3	1.2	15.6	16.5
St. Clara dipl. Danish	15.8	42.1	31.6	0.0	31.6	31.6
St. Clara dipl. Latin	1.4	54.9	56.3	0.0	26.8	29.6
St. Clara facs. Danish	42.1	63.2	57.9	5.3	10.5	10.5
St. Clara facs. Latin	7.0	69.0	71.8	0.0	1.4	2.8
	Linear Support Vector Classification					
Colonia	36.8	47.4	36.8	36.8	42.1	36.8
SDHK Latin	37.5	53.9	53.4	41.3	35.1	34.2
SDHK Swedish	81.0	89.0	88.0	76.2	69.9	69.4
SemEval 2015	26.8	31.4	30.2	28.4	24.7	24.2
St. Clara dipl. Danish	26.3	57.9	36.8	10.5	10.5	10.5
St. Clara dipl. Latin	38.0	47.9	39.4	32.4	19.7	26.8
St. Clara facs. Danish	42.1	31.6	10.5	0.0	10.5	21.1
St. Clara facs. Latin	47.9	49.3	36.6	33.8	15.5	22.5
	Gaussian naive Bayes					
Colonia	31.6	21.1	31.6	26.3	36.8	42.1
SDHK Latin	12.9	37.2	58.3	62.9	-	-
SDHK Swedish	19.8	84.6	92.7	86.9	88.8	-
SemEval 2015	19.7	21.7	39.8	50.9	49.0	43.1
St. Clara dipl. Danish	31.6	26.3	21.1	36.8	47.4	15.8
St. Clara dipl. Latin	16.9	39.4	54.9	54.9	66.2	69.0
St. Clara facs. Danish	26.3	31.6	36.8	36.8	47.4	36.8
St. Clara facs. Latin	53.5	67.6	70.4	63.4	66.2	59.2
	Support Vector Classification with Radial Basis Function					
Colonia	42.1	52.6	42.1	42.1	42.1	42.1
SDHK Latin	45.0	53.4	58.3	48.0	40.1	10.6
SDHK Swedish	88.3	90.3	90.1	80.6	1.3	1.5
SemEval 2015	27.6	30.1	31.6	27.4	10.3	14.3
St. Clara dipl. Danish	26.3	57.9	26.3	21.1	15.8	21.1
St. Clara dipl. Latin	45.1	46.5	50.7	38.0	25.4	19.7
St. Clara facs. Danish	36.8	21.1	31.6	21.1	15.8	26.3
St. Clara facs. Latin	50.7	47.9	45.1	33.8	15.5	25.4
	Multinomial Naive Bayes					
Colonia	26.3	26.3	26.3	26.3	26.3	26.3
SDHK Latin	26.3	26.4	29.9	39.0	39.1	36.7
SDHK Swedish	69.0	69.0	69.0	69.0	69.0	69.0
SemEval 2015	23.6	23.6	23.6	23.6	23.6	23.6
St. Clara dipl. Danish	10.5	21.1	15.8	26.3	21.1	10.5
St. Clara dipl. Latin	19.7	19.7	19.7	22.5	21.1	19.7
St. Clara facs. Danish	26.3	26.3	26.3	26.3	21.1	26.3
St. Clara facs. Latin	25.4	25.4	19.7	25.4	19.7	19.7
	Gaussian Process Regression					
Colonia	21.1	0.0	31.6	0.0	0.0	5.3
SDHK Latin	34.1	35.6	38.9	40.9	28.8	26.6
SDHK Swedish	79.7	75.4	79.3	80.2	3.1	1.5
SemEval 2015	12.8	17.1	15.3	12.3	8.3	6.9
St. Clara dipl. Danish	15.8	21.1	47.4	31.6	26.3	21.1
St. Clara dipl. Latin	23.9	22.5	29.6	25.4	16.9	16.9
St. Clara facs. Danish	21.1	47.4	52.6	21.1	26.3	21.1
St. Clara facs. Latin	54.9	42.3	42.3	31.0	14.1	16.9