Explainable Publication Year Prediction of Eighteenth Century Texts with the BERT Model

Rastas, Iiro

The Association for Computational Linguistics
2022

Rastas , I , Ryan , Y C , Tiihonen , I L I , Qaraei , M , Repo , L , Babbar , R , Mäkelä , E , Tolonen , M & Ginter , F 2022 , Explainable Publication Year Prediction of Eighteenth Century Texts with the BERT Model . in N Tahmasebi , S Montariol , A Kutuzov , S Hengchen , H Dubossarsky & L Borin (eds) , Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change . The Association for þÿComputational Linguistics , Stroudsburg , pp. 68 77 , Workshop on Co
Approaches to Historical Language Change , Dublin , Ireland , 26/05/2022 . < https://aclanthology.org/2022.lchange-1.7.pdf >

http://hdl.handle.net/10138/344533

Explainable Publication Year Prediction of Eighteenth Century Texts with the BERT Model

Iiro Rastas[1], **Yann Ryan**[2], **Iiro Tiihonen**[2], **Mohammadreza Qaraei**[3], **Liina Repo**[1], **Rohit Babbar**[3], **Eetu Mäkelä**[2], **Mikko Tolonen**[2], **Filip Ginter**[1]

[1] TurkuNLP, University of Turku, Finland
[2] University of Helsinki, Finland
[3] Aalto University, Finland

iiro.t.rastas@utu.fi, yann.ryan@helsinki.fi
iiro.tiihonen@helsinki.fi
mohammadreza.mohammadniaqaraei@aalto.fi
tlkrep@utu.fi, rohit.babbar@aalto.fi
eetu.makela@helsinki.fi, mikko.tolonen@helsinki.fi
figint@utu.fi

Abstract

In this paper, we describe a BERT model trained on the Eighteenth Century Collections Online (ECCO) dataset of digitized documents. The ECCO dataset poses unique modelling challenges due to the presence of Optical Character Recognition (OCR) artifacts. We establish the performance of the BERT model on a publication year prediction task against linear baseline models and human judgement, finding the BERT model to be superior to both and able to date the works, on average, with less than 7 years absolute error. We also explore how language change over time affects the model by analyzing the features the model uses for publication year predictions as given by the Integrated Gradients model explanation method.

1 Introduction

Collections of historical language, such as ECCO which comprises over 180,000 titles published in the eighteenth century, are at the focus of a growing interest in the NLP community. The large quantities of raw textual data in these collections, which may cover whole centuries worth of published works, are suitable for language modelling research, a popular and highly relevant topic in NLP. The historical language itself poses new and interesting challenges, especially due to the fact that the collections span over a time frame long enough to be affected by natural language change and evolution. Furthermore, artifacts relating to the technical process – namely the OCR quality – of the works pose a whole new set of challenges rarely met in modern NLP which mostly deals with born-digital texts, for the most part devoid of such artifacts. These new developments in NLP are crucial also for historians and other humanists applying them to new research questions and ways to produce historical evidence.

The transformer model (Vaswani et al., 2017) and especially the BERT (Devlin et al., 2019) bidirectional encoder based on the transformer, form the foundation of present-day practical NLP research and are naturally also applied in the historical language domain. BERT models have already been trained with various historical data sets and languages, including at least English, German, French, Latin and classical Chinese (Ehrmann et al., 2021; Yu and Wang, 2020; Labusch et al., 2019; Bamman and Burns, 2020). The range of tasks to which it has been used in the domain is already diverse, covering at least named entity recognition, construction of word embeddings, event detection, stance detection, word sense disambiguation and the study of the animacy of target expressions (Hamdi et al., 2021; Sims et al., 2019; Coll Ardanuy et al., 2020; Hosseini et al., 2021; Beelen et al., 2021). Issues particular to the historical language domain have also produced new challenges for BERT appliers to adress, like the effect of OCR quality (Jiang et al., 2021).

In this work, we will follow two directions. Firstly, we set out to train from scratch and release a dedicated BERT model specifically on and for the ECCO dataset. Then, we establish whether such a targeted BERT model provides an advantage over other existing historical English BERT models, or even the modern English BERT. To this end we pursue a benchmark task whereby the model is trained to predict the year of publication based on the text itself. We find that the model performs much better on this task than we intuitively expected, and therefore we carry out and report on a more extensive analysis of the task including a comparison to hu-

Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 68 - 77
May 26-27, 2022 ©2022 Association for Computational Linguistics

man performance, and provide aggregated feature attributions to the BERT model predictions using the Integrated Gradients model explanation method of Sundararajan et al. (2017).

2 Data

ECCO, or Eighteenth Century Collections Online, is a set of digitized documents claimed by its publisher Gale to "contain every significant English-language and foreign-language title printed in the United Kingdom between the years 1701 and 1800" (Gale). In truth however, ECCO is a growing collection. Currently comprising the initial ECCO1 set of around 135,000 documents published in 2003 and some 47,000 further titles added as ECCO2 in 2009, the collection has recently been evaluated as containing about 54% of the works printed in the United Kingdom in the eighteenth century, and known to remain to us through time. Thus, while not complete and at points biased, it is certainly an impressive resource for eighteenth-century scholars as well as, for example, historical linguists (Tolonen et al., 2021).

For the purposes of this work, it is additionally useful to know the following information about ECCO. First, ECCO is temporally skewed toward the end of the eighteenth century, with many more works being published particularly in the final two decades of the century than in earlier ones. Second, while some non-English works are included in the collection, 94% of the documents in it are in English (the other languages with more than 1% representation are French, with 2.7% and Latin with 2.5%). Third, the print quality and thus OCR quality of the documents in ECCO correlates both with their format (pamphlet vs. book) as well as publication date, with more recent publications having a significantly better average OCR quality. Further, OCR quality also differs between ECCO1 and ECCO2, which were scanned and OCR'd using different processes. Finally, there may often be multiple editions of a single work within ECCO, and while they have been printed in the eighteenth century, they may well have originated from e.g. antiquity. Further, when the year of publication of a work has not been printed on its title page, the year has often been estimated. On the level of the whole ECCO data, this manifests itself as frequency spikes on every round fifth, tenth and fiftieth year. (Tolonen et al., 2021)

For the purposes of the year regression experiments in this work, we have dealt with the last two problems by limiting the subset of ECCO we are experimenting on to only those where the year of publication is certain, as well as only to the first editions of works that first appear in the eighteenth century. The size of this subset is approximately 40,000 documents.

3 Methods

In this section, we describe the models and methods used in this work: the pre-trained BERT model, the BERT-based year regression and feature attribution, and finally the linear baseline.

3.1 BERT pre-training

BERT model pre-training on the ECCO dataset is very similar to pre-training on any other dataset, with the structure of the dataset and the OCR noise present requiring some consideration. The sub-word vocabulary of size 50,000 is induced in the standard manner on a random sample of the dataset. For the BERT training objective which includes the next sentence prediction task, the training examples are constructed from pairs of text segments. Here each text segment is a continuous piece of text drawn from a single block of text in ECCO, where each such block of text is delimited by an empty line and corresponds to one page or one paragraph, depending on the format of the underlying work. We make an attempt to respect sentence boundaries when forming the training text segments using a simple regular expression, while keeping each segment between 128-384 tokens long, the pair subsequently trimmed to the model's maximum sequence length of 512 input tokens (sub-words and special tokens). Unlike for all other experiments, the entire ECCO dataset is used for BERT pre-training. The trained model is equal in size to the BERT Base models of Devlin et al. (2019). The final model was pre-trained for 1 million steps, with an effective batch size of 768, and learning rate 1×10^{-4}.

3.2 BERT-based year regression

As the regression model, we employ a simple linear regression layer on top of the pre-trained BERT model, as illustrated in Figure 1. The model is trained using the mean square error (MSE) objective. To ensure good model performance, the target values are z-transformed, $y' = \frac{y-\mu}{\sigma}$ where μ and σ are the mean and standard deviation of the publication years of the training set examples. The z-transformed years are centered on zero with a unit standard deviation. While this is a trivial linear transformation, it is crucial in model training: the randomly initialized output regression layer initially predicts values around 0 and a large number of training steps are needed to reach the target range of 1701–1800. During these training steps, the gradients are propagated also into the BERT model, and the combined effect turns out to be highly detrimental for the model.

The documents in our dataset, full books for the most part, are naturally considerably longer than the maximum sequence length of 512 subwords for the BERT model. We therefore split each document into a number of chunks of up to 512 sub-words in length, and subsequently average the predictions to obtain a single, document-level prediction.

Even though the ECCO works (books and pamphlets) themselves are long relative to the maximum sequence length of the model, we originally restricted the textual segments used as inputs to within a single textual block (page/paragraph) of the source document, so as to match the data on which the model was pre-trained. Many of these are relatively short, due both to the layout of the works and OCR artifacts. Unsurprisingly, though, we found during development that the prediction performance is best on long textual segments near the 512 sub-words limit. Therefore, we altered the example generation strategy and concatenated what would originally be several independent examples into a single long sequence separated by the `[SEP]` BERT control tokens. This way the model can be trained and evaluated exclusively on 512 sub-word long segments with the exception of document-ending segments and the rare cases where the entire document is shorter than 512 subwords.

3.3 Feature attributions

There are numerous methods for calculating *feature attributions*, i.e. the assignment of importance to input features with respect to the prediction made by the model. In this work, we apply the Integrated Gradients (IG) method of Sundararajan et al. (2017) to obtain attributions for the BERT-based regressor predictions. IG is a popular method specifically targeting differentiable models, assigning attributions to individual parameters of the model. In the context of BERT, the attributions would typically be calculated with respect to the input sub-word embeddings, in turn providing attributions on the level of sub-words in the input sequence. In short, the IG method defines the attribution as the integral of the gradient of the model output w.r.t. the parameter of interest, integrated on a path interpolating between a "blank" reference input sequence and the actual input sequence. This is in practice implemented by evaluating the model in N steps (here we set $N = 50$) between the reference and actual input.

In image processing, the reference input would typically be e.g. an empty image, or a white noise image. In the context of BERT, we can use the sequence `[CLS] [PAD] [PAD] ... [PAD] [SEP]`, where `[CLS]` and `[SEP]` are the special separation tokens in BERT input, and `[PAD]` is the padding token. This reference sequence has same length as the actual input and the interpolation is carried out on the input token embedding vectors.

The attribution value of each input sub-word is the sum of the scalar attributions across the dimensions of the input embeddings. A positive attribution value signals contribution *towards* the prediction made by the model, while a negative attribution value signals contribution *against* the prediction made by the model. Since the BERT model uses sub-word tokenization, splitting rare words into sub-words, to obtain word-level attributions understandable to the human reader we set the attribution of a word to be the attribution of that of its sub-words which has the highest absolute value. Thus, for instance, if an input word is divided into three sub-words with attributions of $[-0.4, 0.1, 0.21]$, the overall attribution of the word will be -0.4.

3.4 Aggregating attributions

The word attributions provided by the IG method are assigned to individual predictions, i.e. predictions on a maximum of 512 sub-words long text segments. There are therefore two levels on which the attributions may be aggregated. Firstly, relevant features aggregated across all text segments of a single long document such as a book explain the prediction the model gave to that document. And secondly, one might be interested in aggregating relevant features across all books published in a single period (e.g. one decade) so as to gain an understanding of globally relevant features for that period.

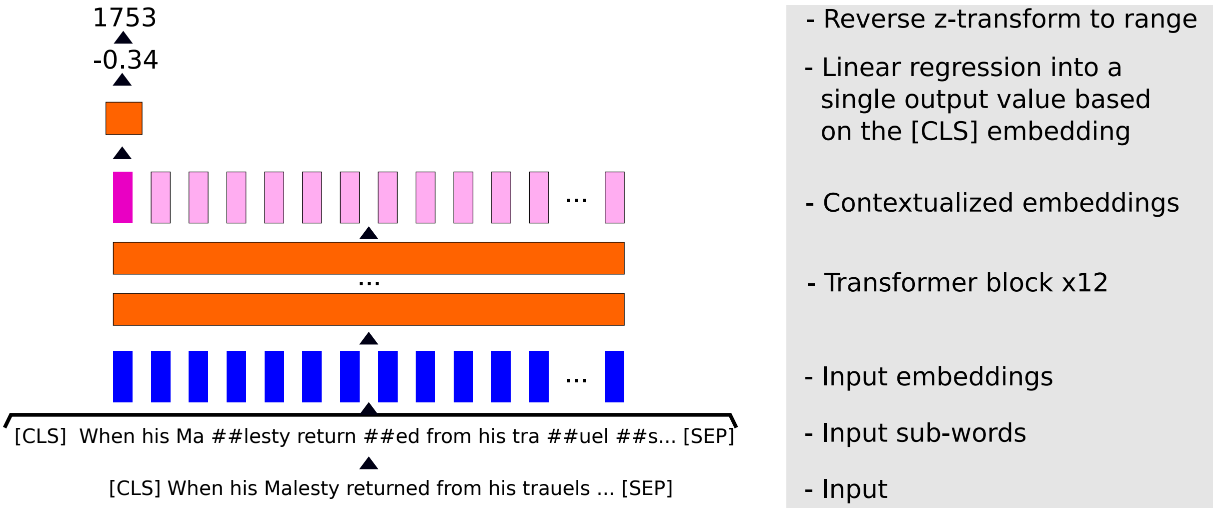

1753
-0.34
...
...
...
[CLS] When his Ma ##lesty return ##ed from his tra ##uel ##s... [SEP]
[CLS] When his Malesty returned from his trauels ... [SEP]
- Reverse z-transform to range
- Linear regression into a single output value based on the [CLS] embedding
- Contextualized embeddings
- Transformer block x12
- Input embeddings
- Input sub-words
- Input

Figure 1: The regression model for a single text segment of BERT maximum sequence length, with OCR errors. Predictions across segments of a single document / book are averaged to give a final document-level prediction.

Model	**MAE**	**MSD**	**STD**
ECCO-BERT	6.32	-1.30	8.84
dbmdz/bert-base-historic-english-cased	7.27	-1.44	10.18
bert-base-cased	7.65	-0.73	10.27
MacBERTh	8.21	-1.35	11.08
Linear regression	11.88	0.26	15.38
Linear classification	12.47	-0.35	20.22

Table 1: Results for fine-tuned BERT models and the linear baseline models. MAE is mean absolute error, MSD is mean signed deviation, and STD is standard deviation, in terms of years.

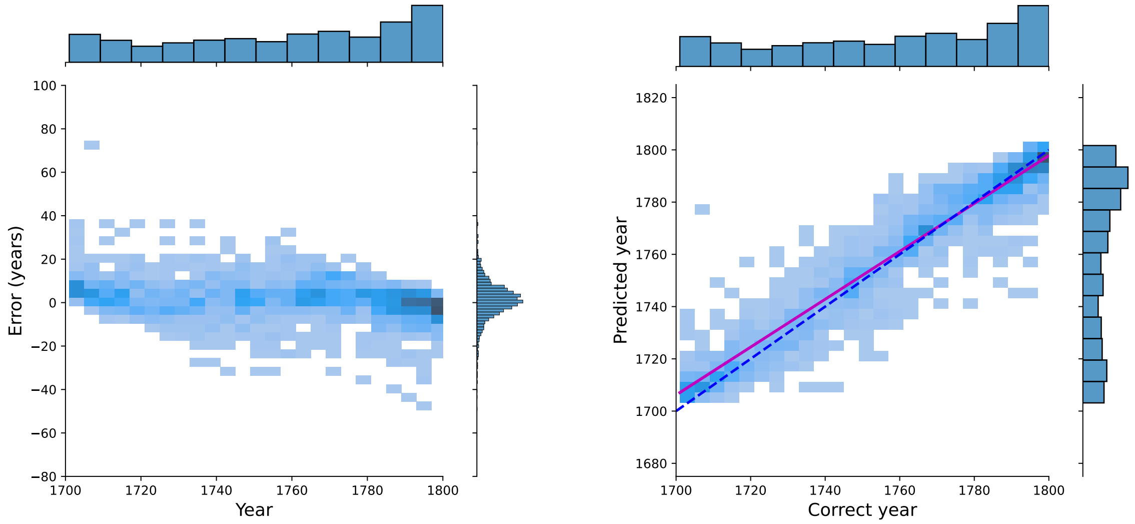
There are many ways to approach this aggregation. In the simplest case, we can take the top features based on the highest attribution values across all text segments. However, this method was found to be prone to noise when the number of segments is large, such as when aggregating on a decade-level. We therefore test two additional methods. The first one counts the number of times each feature appears as a top 10 feature of a segment. To reduce the prevalence of common words, this number is further weighed with its IDF. The other method takes the average attribution value for each feature across all segments. Top features are chosen as those that have the highest average attribution value and appear in the segments more than once. Using these methods, lists of top features for each decade were qualitatively evaluated.

3.5 Linear baseline

We use a standard linear model as the baseline method, as it also allows us to compare the feature attributions, which are simple to extract from a trained linear model. As the first baseline to evaluate the performance of a linear model using support vector regression on the task of year prediction, we used the solver implemented in the Liblinear package (Fan et al., 2008). As an alternative, we also used a linear model for the direct multiclass prediction (Crammer and Singer, 2001) instead of the surrogate loss in the form of squared error.

4 Results

There are 39,429 ECCO works that have a verified year of publication during the 18th century and that constitute the earliest publication of the given work. All results are reported on the same test set of 1971 randomly selected works, which contain a total of about 225,000 text segments. A development set of the same size was reserved for hyperparameter selection, with the remaining 35,487 documents being used for training.



(a)

(b)

Figure 2: A histogram of the errors (a) as well as a comparison between the actual years and predicted years (b) by the ECCO-BERT model.

4.1 Year regression

In addition to our pre-trained ECCO-BERT, we used three other relevant BERT models pre-trained on either historical or modern English: bert-base-cased[1], dbmdz/bert-base-historic-english-cased[2], and MacBERTh[3]. For each model, a grid search on the development set was performed to find optimal learning rate and number of training steps.

The overall results for the year regression task with the BERT models as well as the linear baseline models are summarized in Table 1. All BERT models can be fine-tuned to perform reasonably well, as the fine-tuning dataset is very large. BERT pre-trained on the ECCO dataset performs slightly better than the other models, possibly due to better fitting the OCR noise unique to the dataset. Overall, the best result of mean absolute error of 6.32 years reflects a surprisingly good performance of the BERT model on the task. To gain more insight into the predictions, the histograms of prediction errors relative to the publication year of the work are presented in Figure 2. These show no strong bias, beyond the natural fact that the publication year of older works is more likely to be overestimated and the publication year of newer works is likely to be underestimated, as the model learned the prediction range.

The impact of OCR quality on the results is worth considering. (Jiang et al., 2021) showed that pre-trained BERT on OCR'd historical books was less robust when used in a domain classification task than one trained on 'clean' text, though in that study fine-tuning significantly improved resilience to noise. Other studies on downstream NLP and language modelling tasks show that OCR quality can have a significant effect, though the extent is heavily dependant on the specific task and extent of the OCR error rate. (van Strien. et al., 2020; Hill and Hengchen, 2019) Here, we found a moderate performance difference between the ECCO1 and ECCO2 subsets of the test set, with ECCO-BERT having a mean absolute error of 6.96 years for ECCO2, but only 5.95 years for ECCO1. This is most likely due to ECCO1 having a stronger relationship between OCR quality and publication year, which could help model predictions. This suggests that a more noise-aware variant of the model, for instance a character-level version, would improve results.

4.2 Linear baseline

Contrary to the BERT model, in which the input is limited to 512 sub-words, in the linear models, the TF-IDF representation can be built over the entire corpus. For building a TF-IDF representation for each document, we used the `TfidfVectorizer` of `sklearn`. We ignored the terms that ap-

[1] https://huggingface.co/bert-base-cased

[2] https://huggingface.co/dbmdz/bert-base-historic-english-cased

[3] https://huggingface.co/emanjavacas/MacBERTh

pear in more than `tf-max` $= 30\%$ or less than `tf-min` $= 1\%$ of the training documents. As the data contains a significant amount of noise, the only preprocessing on token-level is removing the stop words. Furthermore, to prevent information leakage when the year of publication is explicitly stated somewhere in the document, we removed all the numbers from the documents including training and test sets[4]. The hyper-parameters of the linear models including C, `tf-max`, and `tf-min` are chosen using a validation set drawn randomly from 5% of the training data.

The histograms of the errors using linear models are depicted in Figure 3. While it seems that the classification model is more accurate in predicting the distribution of the years, having predictions with large variances leads to worse performance of this model compared to linear regression when metrics such as standard deviation are taken into account.

Overall, as can be expected, the linear baselines performs substantially worse than any of the BERT models, including modern English BERT.

4.3 Qualitative evaluation

Three approaches were used to analyse what information carried by the text tokens the BERT model might be utilising in its predictions. First, we qualitatively evaluated the predictor features of the linear regression model used as the baseline. This evaluation suggests that - even when the model is simple and features easier to interpret - there are multiple elements in the ECCO's tokens that a year predicting model can use. Some like *baptizing* (negative predictor, i.e. signalling an old publication) might relate to shifts in the composition of the ECCO during eighteenth century, others like *soveraign* and *cloath* might be related to temporal variation in spelling. Further likely information sources include language (tokens in Latin and French are prominent among negative predictors) and varying heuristics like the information that is part of the imprint [5]. For example, the term *sixpence* has high positive effect, and sixpence is a very common price printed to the imprint, but price information in ESTC is temporally varying, and mostly missing from the first years of the eighteenth century in contrast to the rest of the century (Tiihonen et al., 2021). In nearly all specific instances there is a high degree of uncertainty about the reason why a given token is or seems to be relevant for year prediction, but put together, the evaluation of the baseline model suggests that there is real information to be utilised among the noise.

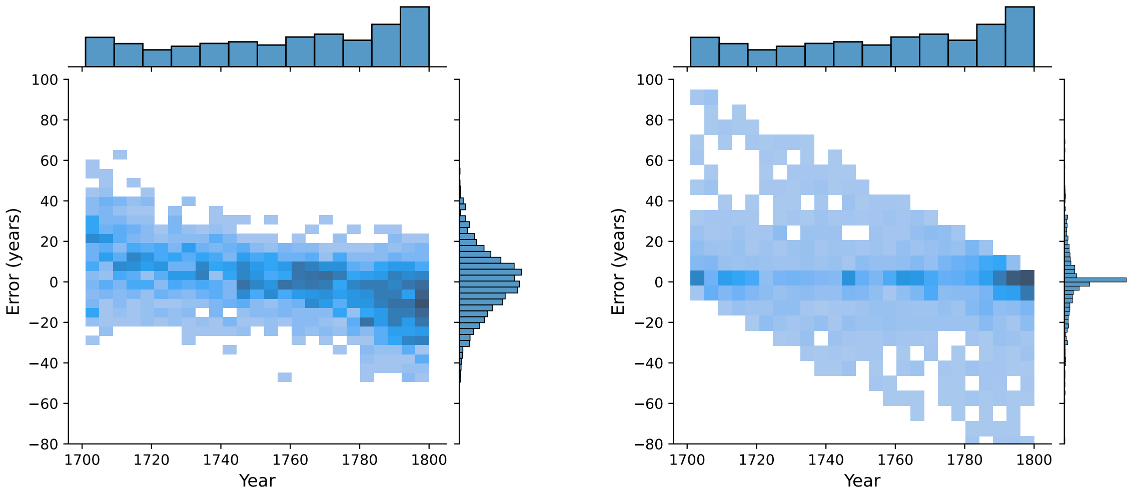
In the second approach, we tried to directly evaluate the predictors relevant for the ECCO-BERT model's predictions by going through a sample of documents and interpreting three sets of predictor tokens for each. Each of these token sets relates to one of the methods of measuring the token's significance as a predictor (see section 3.4), and the motivation was to use these sets of terms to get insight into the way the model utilises information from ECCO to predict the years. In addition to the sources of variation already mentioned, the model seems to capture some very context specific terms relevant for prediction. A telling example is a work[6] on the French Revolution from 1797, that the model predicted as being published in 1794. Among the top predictor tokens for this document were *French*, *Revolution* and *Jacobins* from the second set of tokens, but also *1792*, *I792* and *r792* from the third.[7] In the third approach, the three methods were used to produce three token sets of potentially relevant predictors of the ECCO-BERT model for each decade (the approach discussed in section 3.4) of the eighteenth century. Some of the temporal development of token sets two and three might be related to significant conceptual developments that occurred during eighteenth century. For example, the term *publick* is part of the token set 2 in 1750's, and *Public* in 1790's. The emergence of the notion of a public sphere (for definition, see for example (Barker, 2004)) and the term *public(k)* during the eighteenth century are major questions both in intellectual history and political theory. The transformation from publick to public is an example of known orthographic shift where the letter *k* of words ending with *ck* drops out (Baron, 2011). Both the appearance of the term as a potentially relevant predictor for specific decades and the vari-

[4] Although the numbers are removed from the documents for the experiments with the linear models, we observed in practice that having numbers in the documents may not affect the results significantly, where the MSE metric for linear regression is 232.18 when the numbers are present and 236.71 for the other case.

[5] The text, usually at the bottom of the title page, giving the details of the book's producers a well as information on price and place of publication

[6] ESTC citation number T64288.

[7] Note that as the works are split into a large number of text segments, on average over 100 per work, whose predictions are averaged, a single segment with the correct year picked up by the model does not uniquely decide the result.



(a)

(b)

Figure 3: A histogram of the errors for linear regression (a) linear classification (b).

ation in its spelling are interesting phenomenon from the humanities perspective.

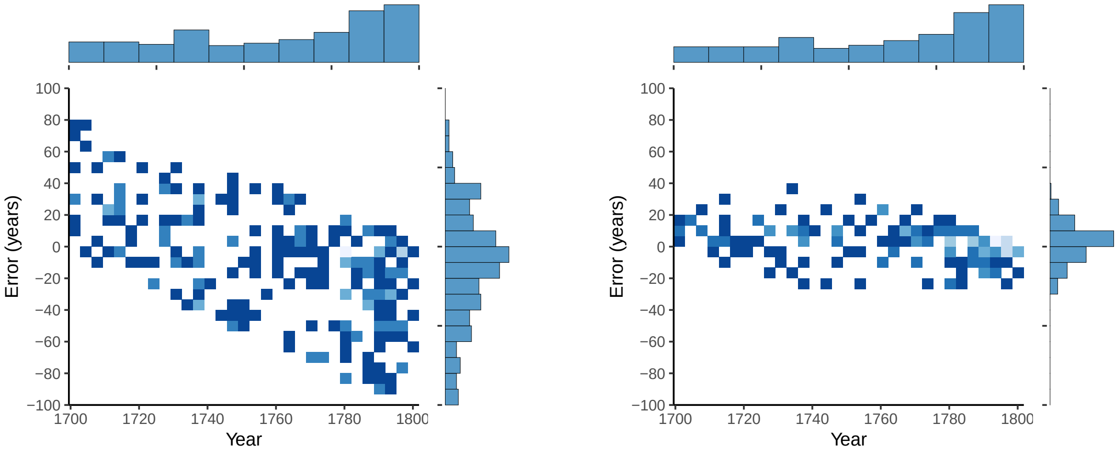
4.4 Manual annotation

As a point of comparison, a set of human-annotated predictions was also produced. Four human annotators were provided with a set of 512-token documents and asked to predict the year of publication. In addition, they were invited to label the features within each document which they determined had been most useful in making a given decision. The annotators all had some level of familiarity or expertise with early modern texts, nevertheless this is still best considered as an initial exploratory study, rather than a fully authoritative experiment. Most importantly, the human annotators did not study each work in its entirety, unlike the models.

In total 277 human predictions were gathered, from 167 distinct document snippets. Human annotators fared much worse than the BERT model predictions for the same set of documents, with a mean absolute error of 30 (27.59 if the average for multiple guesses for the same document is taken) compared with 8.73 for the model (Figure 4a, Figure 4b). Human annotators tended to over-estimate the publication year (Figure 4a). The average errors were higher for documents published towards the end of the century, though this may be partially explained by the fact, noted in section 2, that the labelled data is also biased with more occurrences towards the end (reflecting the distribution of the full dataset).

When comparing the predictive features of the model with those given by the human annotators, categorical or thematic overlap was observed. In many cases the human annotators found it difficult to articulate reasons or pick out specific words to describe how their decision was made, but where they did, it was a mixture of recognition of spelling variations (for example, the additional e in newes, or k in publick), judgements on OCR quality – which improves significantly for documents published later in the century due to improvements in print quality and subsequent digitisation – and historical evidence, for example the mention of a known historical figure or event making it possible to give the earliest possible publication date with certainty, at least. Historical clues ranged from anything from mentions of specific events (such as the resignation of Lord North which took place in 1782) to less obvious historical clues such as the mention in a document of 'hot-house grapes', a growing technique more likely to have been used at the end of the century.

By most accounts, spelling variation in English printed works had already levelled off by 1700 (Baron et al., 2009) meaning that in theory the usefulness of orthographic change as a feature is minimal. However *some* variation is still found, particularly in the earlier part of the eighteenth century, which may account for the fact that human annotators were moderately better at predicting earlier works than later. As expected, a key clue in predicting earlier dates are OCR errors and long-s words transliterated as f. This may also be part of the reason why average errors were highest towards



(a)

(b)

Figure 4: Comparison of human annotations (a) and model predictions (b) for the same set of 167 document snippets.

the very end of the period: the use of the long s declined rapidly by 1800 and so is a less useful clue for dating a document.

The annotators reported that the task was difficult, particularly when judging a year of publication from what was usually a snippet from a much larger text. While reprints in ECCO have been removed as described in section 2, the partial re-use of text is common, for example in miscellanies, anthologies, and collected works. One consequence of this is that typical humanistic features of text such as style of writing were not always helpful in a decision about year of publication. To give one example, an incorrectly-labelled (by a human) annotation included part of a poem written by John Sheffield, Duke of Buckingham who died in 1721, but was actually in this case from a collected works published in 1780 and thus labelled with the later date in the task. Overall, spelling, OCR artifacts, and typographical changes were more useful as predictive features.

The annotation task, then, was valuable firstly as a way to understand the differences between the ways a machine model and human annotator might use features to predict years. Secondly, it shed some light on the way those with domain expertise might judge the year of publication of a particular work based on its text abstracted from the material context in which it was found. From a humanistic point of view, the task highlights the fact that human judgement of publication dates is very unreliable when dealing with extracts from larger texts, and presumably relies to a great extent on contextual information, for example font, paper, and the condition of a particular book, rather than its content.

5 Conclusions

The contributions of the paper are two-fold. Firstly, we pre-trained and openly distribute a BERT model specifically focusing on the historical English language in the Eighteenth Century Collections Online (ECCO) that is widely used in the humanities. To benchmark the model and gain understanding of its performance on historical English, we use the task of publication year prediction, in other words given the text, the task is to regress its publication year.

Our findings and analysis of the model's performance on this task then form the second contribution of the paper. We establish that the accuracy with which the model is able to predict the year of publication is well above our baseline models on full documents and also well above human performance on text snippets.

We also carried out an initial qualitative analysis of predictive features, both for our simple linear baselines and for the BERT models. We observe a degree of a useful signal among these features, intuitively understandable to a human, demonstrating the applicability of model explanation techniques also to the complex BERT model. Nevertheless, it is clear that numerous challenges still remain.

This initial study has several natural future work directions. Firstly, a further, more detailed analysis of the predictive features, and therefore of the model's predictions is clearly called for. Secondly, a more detailed comparison between human and model decisions will be carried out. And finally, as we have not specifically taken into account the OCR noise when pre-training the BERT models, more noise-aware variants of the transformer model, e.g. character-based models, will be tested on the ECCO data. The ECCO-BERT model is freely available as `TurkuNLP/eccobert-base-cased-v1` in the Hugging Face model repository.

Acknowledgements

The research was supported by the Academy of Finland under the project *High Performance Computing for the Detection and Analysis of Historical Discourses*. Computational resources were provided by CSC – IT Center for Science.

References

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology. *CoRR*, abs/2009.10053.

Chris Barker. 2004. *The SAGE dictionary of cultural studies*. Sage Publications Ltd.

Alistair Baron. 2011. *Dealing with spelling variation in Early Modern English texts*. Ph.D. thesis, Lancaster University.

Alistair Baron, Paul Rayson, and Dawn Archer. 2009. Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies*, 20:41–67.

Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 2021. When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2751–2761, Online. Association for Computational Linguistics.

Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. 2020. Living machines: A study of atypical animacy. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4534–4545, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Koby Crammer and Yoram Singer. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification on historical documents: A survey. *ArXiv*, abs/2109.11406.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Gale. Eighteenth Century Collections Online.

Ahmed Hamdi, Elvys Linhares Pontes, Emanuela Boros, Thi-Tuyet-Hai Nguyen, Günter Hackl, Jose G. Moreno, and Antoine Doucet. 2021. A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Mark J Hill and Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4):825–843.

Kasra Hosseini, Kaspar Beelen, Giovanni Colavizza, and Mariona Coll Ardanuy. 2021. Neural language models for nineteenth-century english.

Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C. Dubnicek, Ted Underwood, and J. Stephen Downie. 2021. Impact of ocr quality on bert embeddings in the domain classification of book excerpts. *CEUR Workshop Proceedings*, 2989:266–279. Publisher Copyright: © 2021 Copyright for this paper by its authors.; 2021 Conference on Computational Humanities Research, CHR 2021 ; Conference date: 17-11-2021 Through 19-11-2021.

Kai Labusch, Clemens Neudecker, and David Zellhöfer. 2019. Bert for named entity recognition in contemporary and historic german. In *KONVENS*.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary Event Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Iiro Tiihonen, Mikko Tolonen, and Leo Lahti. 2021. Probabilistic analysis of early modern british book prices. volume 2989 of *CEUR Workshop Proceedings*, pages 39–48. CEUR-WS.org.

Mikko Tolonen, Eetu Mäkelä, Ali Ijaz, and Leo Lahti. 2021. Corpus linguistics and eighteenth century collections online (ecco). *Research in Corpus Linguistics*, 9(1):19–34.

Daniel van Strien., Kaspar Beelen., Mariona Ardanuy., Kasra Hosseini., Barbara McGillivray., and Giovanni Colavizza. 2020. Assessing the impact of ocr quality on downstream nlp tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH,*, pages 484–496. INSTICC, SciTePress.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Peng Yu and Xin Wang. 2020. Bert-based named entity recognition in chinese twenty-four histories. In *Web Information Systems and Applications*, pages 289–301, Cham. Springer International Publishing.