Dating Documents: A Domain Independent Approach to Predict Year of Authorship

Anonymous EMNLP submission

Abstract

We present two classes of computational models to date documents based on linguistic markers. The first class infers global usage patterns of neologisms over time to assign dates to texts, providing insights into temporal locality of word usage. Our second class of models are based on a neural based approach that exploits deeper linguistic cues. We demonstrate that our models generalize across various domains like News, Fiction and Non Fiction over a span of 150 years. Finally, we apply our model to books written by authors over their literary careers which yields insights into the temporal patterns of language used by authors.

1 Introduction

Determining when a document is written is an important task in historical linguistics and temporal information retrieval. For instance, several works attempt to date historical biblical texts like the *The Book of Isiah* (Rooker, 1996; Ehrensvärd, 1997; Hurvitz, 2000; Young and Rezetko, 2016) or ancient texts like *Beowoulf* (Chase, 1997). Likewise, in the field of information retrieval, establishing the dates of documents is an important pre-requisite to returning temporally relevant documents and provides important information for a large number of search tasks (Ostroumova Prokhorenkova et al., 2016; Efron, 2013).

Most efforts to automatically date texts adopt a learning based approach and rely on several linguistic features that are time-relevant (Garcia-Fernandez et al., 2011; Jatowt and Tanaka, 2012; Zampieri et al., 2015; Ostroumova Prokhorenkova et al., 2016). Such time-relevant features include neologisms/archaisms, political events, spelling variations, and the presence of named entities as well as external knowledge bases. Moreover, these approaches all learn models using a training corpus specific to the domains they are evaluated on (for example, primarily News articles).

Differing from previous work, we investigate the problem of dating documents through the *lens of language evolution* over time [1]. We propose models that effectively generalize across a variety of domains (Fiction, Non-Fiction, and News) without further tuning.

First, we develop models that leverage only cumulative usage patterns of neologisms to date documents. These models provide insight into the temporal locality of neologisms by authors, namely that documents written at time t tend to use neologisms invented shortly before t. They perform competitively across domains when compared to other methods that use more fine-grained linguistic cues.

Second, we propose a novel neural model by modeling the problem as a classification task. We observe that unlike standard classification tasks, our classes (labels) have a linear sequential structure which can be effectively exploited to learn a model. Therefore we propose a novel regularizer – an "autocorrelation regularizer", to encourage our model to respect this linear sequential structure in labels and demonstrate competitive performance. We show that using the regularizer improves the performance of our neural model by atleast 2 points in terms of error rate reduction.

Finally, we apply our model to perform a preliminary analysis of broad literary styles of authors over time, where our analysis suggests that authors tend to evolve their language use slowly over time.

To summarize. our contributions are as follows:

[1] Our tool will be released online.

- **Neologism based models**: We propose several models that infer statistical patterns in the usage of neologisms to date documents. We demonstrate that neologism-based models that uses ~ 200 features achieves a performance within 5 units of error (21.58 on NonFiction) over our best Naive Bayes model (18.25 on NonFiction) which uses more than $200K$ features.
- **Autocorrelation regularizer**: We propose a novel regularizer that enables a model to effectively leverage sequential structure of labels (years) for the purpose of dating documents. We empirically demonstrate the efficacy of the regularizer by providing evidence that the regularizer consistently drops the mean error of our neural model on our task by at-least 2 points.
- **Applications**: Our methods demonstrate competitive performance across a wide variety of domains (like fiction, non-fiction and news). Finally, we apply our methods to analyze authors over their literary career, revealing insights into their temporal patterns of language use.

2 Datasets

Here, we describe data-sets over which we comprehensively evaluate our methods:

- **NYTimes**: We consider a random sample of 10000 leading paragraphs of NEW YORK TIMES articles from the range $1850 - 2005$.
- **Corpus of Historical American English (COHA)**: We consider a random sample of 10000 articles from each genre, namely FICTION, NONFICTION and NEWS from the COHA corpus (Davies, 2002).

3 Baselines

We introduce two baseline methods to evaluate against on the task of dating texts.

BOOKPROP We estimate the probability of a document written in a given year y, by computing the fraction of books written in year y over all books written in the time period under consideration. We estimate the number of books in English written in year y, as the number of distinct books the word `the` was mentioned in a given year y as per Google Book NGrams (Michel et al., 2011) data. Given a document to date, a random sample drawn from this distribution is then taken as the predicted estimate of the date of the document. A limitation of BOOKPROP is that it does not model language.

NEO A more sophisticated approach to assigning dates to documents is based on the following observation: If we observe a word which first came into popular usage in a year y, then the document is very likely written after year y. A simple model based on this hypothesis is to output the year of the most recent word found in the document. For example, in Figure 1, NEO estimates the date of a document to be 1958, since it is clear that `Sputnik`, the most recent word used in the document, sprung into popular use in 1958.

We estimate the year in which a word came into popular usage MR(w) from Google Book NGrams as follows: (a) Compute the cumulative usage of a word w through every year in the Google Book NGrams. (b) Compute the first year in which the cumulative usage of the word w exceeds a small fraction α of the total cumulative usage.[2] As an example, our method estimates MR(`Obama`) $=$ 2007 while the year of actual first usage is 2006.[3].

While NEO serves as a strong baseline, there are two limitations of this method: (a) The document could be written long after the time period corresponding to the most recent word observed in the document. (b) It ignores evidence of other words seen in the document and bases its decision on the occurrence of a single word.

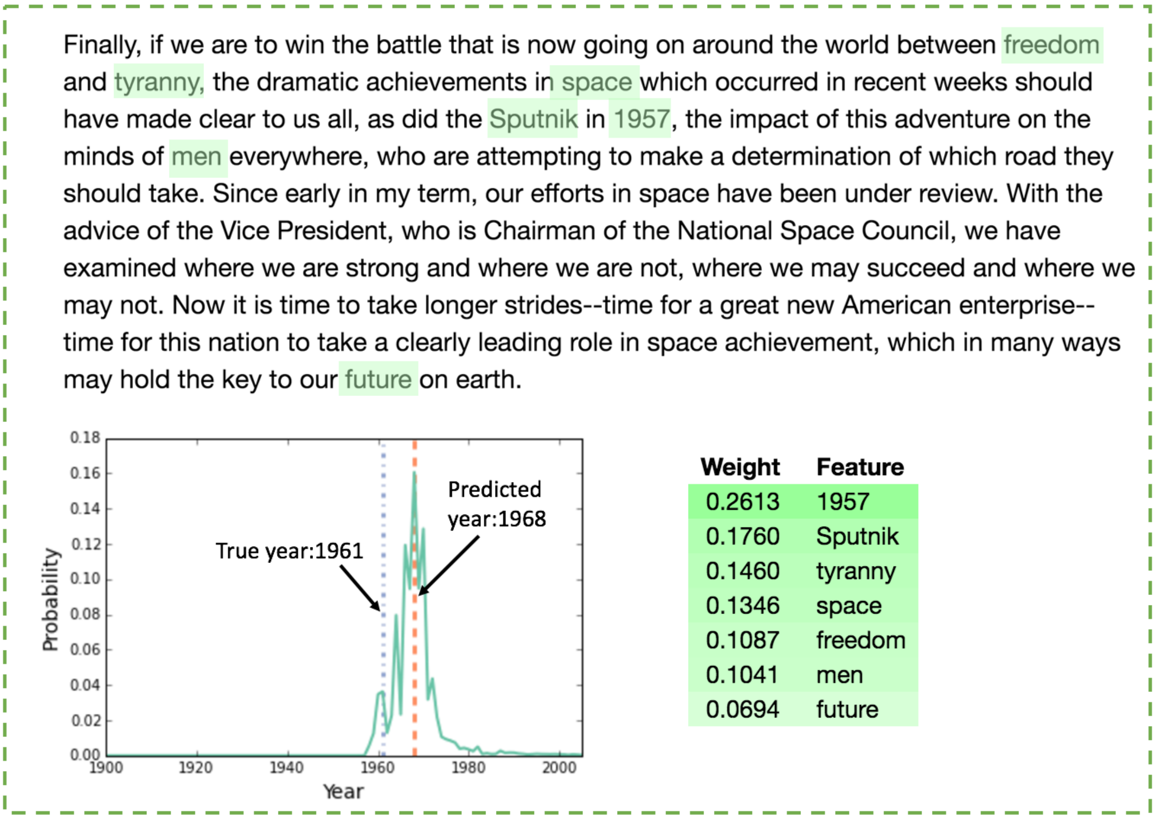
Figure 2a illustrates these drawbacks. First, note that only one word `vinaya` with an MR(w) in the 1990's was observed in this document. NEO is easily misled by this single erroneous estimation of MR(`vinaya`) and estimates the date of this document to be 1992. It ignores other evidence that suggests that the document was written after 1940, but is unlikely to be written in the 1990's since words with MR(w) a few decades before 1990's are not observed at all.

4 Methods

Here we propose several count-based models as well a neural model for dating texts.

[2] We set $\alpha = 1/250$ empirically.

[3] Actual first usage are obtained from http://www.etymonline.com/. See Table 3 in supplementary material for more examples



Finally, if we are to win the battle that is now going on around the world between freedom and tyranny, the dramatic achievements in space which occurred in recent weeks should have made clear to us all, as did the Sputnik in 1957, the impact of this adventure on the minds of men everywhere, who are attempting to make a determination of which road they should take. Since early in my term, our efforts in space have been under review. With the advice of the Vice President, who is Chairman of the National Space Council, we have examined where we are strong and where we are not, where we may succeed and where we may not. Now it is time to take longer strides--time for a great new American enterprise--time for this nation to take a clearly leading role in space achievement, which in many ways may hold the key to our future on earth.

Weight	Feature
0.2613	1957
0.1760	Sputnik
0.1460	tyranny
0.1346	space
0.1087	freedom
0.1041	men
0.0694	future

Figure 1: Sample output predictions of our model (NB) on a portion of a speech given by President John.F.Kennedy in 1961. Note that our model outputs a probability distribution over years with a MAP estimate of year 1968. Note also that the model was trained *only using* Google Book NGrams and not on the domain it is evaluated on. Finally observe that words like `1957, Sputnik, tyranny and space` were most influential in this prediction thus providing insight into linguistic patterns the model has captured only from Google Book NGrams. Our model is generic and can be applied to multiple domains like Fiction, News or Non Fiction.

4.1 Count Based Models

4.1.1 NEO-PROB

We now describe a probabilistic model that effectively uses new words incorporated into popular usage to estimate when the document was written. In particular, our model computes the likelihood of observing a set of words that came into popular usage after year x given the document was written in year y to estimate when the document was written. Our method has two key steps:

1. **Ensemble Model Construction**: We construct an ensemble of probabilistic models where model M_i outputs $\mathsf{P}(y|X_i)$ and X_i is a discrete random variable counting the words observed in a document which came into popular usage after year i.
2. **Combining Ensemble Predictions**: Each model M_i outputs $\mathsf{P}(y|X_i)$, so we investigate multiple methods to combine predictions from individual models.

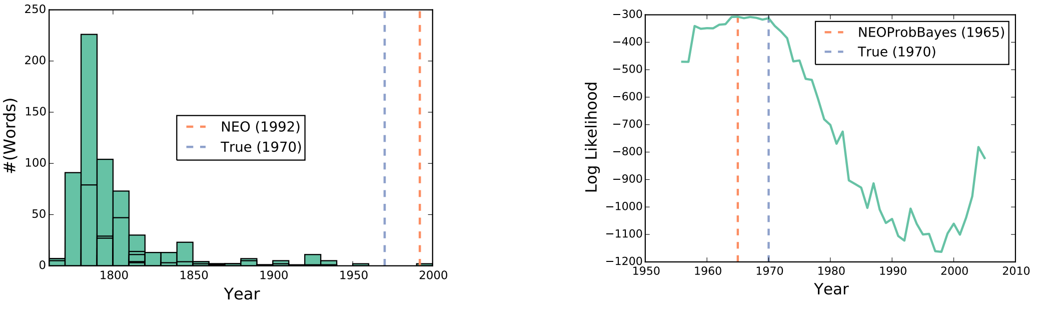
Ensemble Model Construction Let $\mathsf{F}(o, n)$ be the probability of observing a word invented after year o in year n, where $n > o$. For every year pair (o, n), we estimate $\mathsf{F}(o, n)$ from the Google Books NGrams Corpus by computing the fraction of words with $\mathrm{MR}(w) > o$ in the Google Book NGrams of year n.

Given a text T of length N, let $N'(i)$ denote a realization of X_i in T. Each model M_i models the probability of T written in year y based on X_i as follows:

$$\mathsf{P}(y|X_i) \propto \begin{cases} \mathsf{P}(X_i|y)\mathsf{P}(y), & \text{if } i < y \\ 0, & \text{otherwise} \end{cases}$$

$\mathsf{P}(X_i|y)$ follows a Binomial Distribution with success probability $\mathsf{F}(i, y)$. We assume the prior $\mathsf{P}(y)$ to be uniform.

Combining Ensemble Predictions Each model M_i computes a probability distribution over years, namely $\mathsf{P}(y|X_i)$. We now describe three methods to combine these individual model predictions to output a final prediction:



(a) NEO

(b) NEOProbBayes

Figure 2: Figure illustrating NEO and NEOPROBBAYES on the same document which was written in 1970. NEO is easily misled by outlier words and predicts the date to be 1992 ignoring other evidence like counts of other words. NEOPROBBAYES, in contrast incorporates all observed evidence to estimate more accurately that the document was written in 1965.

1. **NEOProbMean**: For each model M_i we compute its MAP prediction $p_i = \arg\max_y (\mathsf{P}(y|X_i))$. The mean of these individual model predictions is the predicted year of authorship.

2. **NEOProbMedian**: We output the median of individual MAP predictions as the predicted year of authorship.

3. **NEOProbBayes**: We use a Bayesian scheme to incorporate all the observed evidence as follows: Let $\boldsymbol{X} = \{X_i \text{ for each year } i\}$. Specifically we compute the following:

$$\begin{aligned} \mathsf{P}(y|\boldsymbol{X}) &\propto \mathsf{P}(\boldsymbol{X}|y)\mathsf{P}(y) \\ &= \left(\prod_i \mathsf{P}(X_i|y)\right)\mathsf{P}(y) \end{aligned}$$

   where we make the *Naive Bayes assumption* that each X_i is independent of any other X_j, when conditioned on the year y. We output the MAP estimate of $\mathsf{P}(y|\boldsymbol{X})$ as the final prediction.

Figure 2b shows this approach for a document and also contrasts it with the baseline NEO. Observe how NEOPROBBAYES enables a more accurate prediction by incorporating observed evidence ignored by NEO.

4.1.2 NAIVEBAYES

We also investigate using a simple, standard Multinomial Naive Bayes classifier learned using Google Book NGrams to date the year a document was written. We use unigram bag-of-words (we restrict our vocabulary to $200K$ tokens and discard out-of-vocabulary words) features and Laplace smoothing. It is worth noting that Naive Bayes uses $200K$ features which are orders of magnitude higher than **NEO-Prob** approaches.

4.2 Neural Model: NEURALDATE

Here, we propose a neural model, NEURALDATE, to date texts. Our model operates on short sequences of words (n-grams). Our model outputs a probability distribution over years, $\mathsf{P}(y|\boldsymbol{x}_i)$ for each ngram $\boldsymbol{x}_i$ in the document $\boldsymbol{D}$,

In order to date a document $\boldsymbol{D}$, we use the model to compute $\mathsf{P}(y|\boldsymbol{x}_i)$ for each n-gram in $\boldsymbol{D}$. We then compute $\mathsf{P}(y|\boldsymbol{D})$ to be the mean of these individual probability distributions. Finally we use the MAP estimate of $\mathsf{P}(y|\boldsymbol{D})$ as our point estimate of the year.

We configure our model to use an embedding of 200 dimensions, hidden states $\boldsymbol{h}_N^l$ of 256 dimensions, and 2 layers. All recurrent matrices are initialized with Xavier initialization (Glorot and Bengio, 2010), all embeddings from a uniform distribution in $[-\sqrt{3}, -\sqrt{3}]$, $\boldsymbol{W}_{\text{out}}$ from Gaussian initialization with $\sigma = 1/|Y|$ truncated at 2σ, and all bias from zero initialization. we use the ADAM optimizer (Kingma and Ba, 2014) with a learning rate of $\eta = 0.001$.

4.2.1 Autocorrelation Regularizer

The model described above does not explicitly leverage structure of the label space, namely temporal structure (linear sequential structure). Observe the high variance in probability scores around the mode in Figure 3. Therefore, for a given n-gram $\boldsymbol{x}_i$ it would be preferable to learn model parameters such that $\mathsf{P}(y|\boldsymbol{x}_i)$ is "smooth" around any given label. This captures the insight that classes (years) in the neighborhood of a label l should be assigned probabilities similar to that assigned to l.

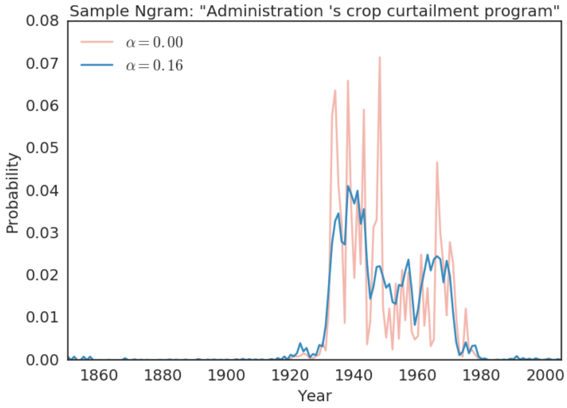


Figure 3: Predicted distribution over years for given 5-gram (shown above the chart), with and without autocorrelation regularization. Note that when $\alpha = 0$, the regularizer is disabled and the output probability distribution is very noisy and neighboring values have large variance. In contrast, when the regularizer is properly enabled ($\alpha = 0.16$), observe how the output probability distribution is much smoother and neighboring probability values are more similar.

We can formalize this notion of smoothness as follows: Let p_l be the probability assigned to label l. Given a neighborhood k, let $\boldsymbol{d}$ be the vector of first order differences: $p_i - p_{i+1}$ for $i \in [l - k \cdots l + k]$. We define the distribution to be L-smooth at l around neighborhood k if

$$\omega(\boldsymbol{d}) = \frac{\sigma(\boldsymbol{d})}{\text{mean}(|\boldsymbol{d}|)} \leq L \,,$$

for some small constant $L > 0$, where smaller values of L indicate smoother distributions.

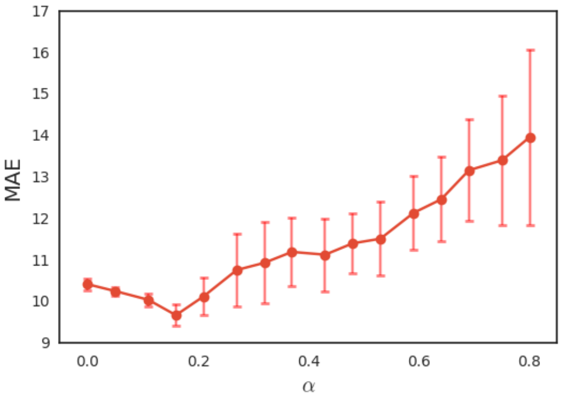


Figure 4: MAE from cross validation for candidates of α. The means and the standard deviations over 20 independent runs are shown. When the model is properly regularized ($\alpha = 0.16$), observe the improvements over model without regularization ($\alpha = 0$). Also note that when the model is over-regularized ($\alpha > 0.25$), the performance is worse and demonstrates larger variance.

We therefore propose to add the following cost to the original cost function:

$$\Omega(\boldsymbol{\theta}; \boldsymbol{X}^{\text{train}}) = \sum_j \omega\left(\boldsymbol{d}_j\right) \,,$$

where $\boldsymbol{d}_j$ are differences between predicted probabilities for neighboring years for example j.

In summary, the final loss function including this regularization is

$$J(\boldsymbol{\theta}; \boldsymbol{X}^{\text{train}}, \boldsymbol{y}^{\text{train}}) + \alpha\Omega(\boldsymbol{\theta}; \boldsymbol{X}^{\text{train}}) \,,$$

where α is a hyper-parameter.

Figure 3 shows the effect of incorporating label smoothness constraints in the cost function for a sample n-gram. Note that incorporating the temporal structure of labels in the cost function produces markedly smoother and realistic distributions than a model not exploiting label structure.

To investigate the effect of α, we measure the MAE (Mean Absolute Error) over n-grams and use cross-validation by selecting α from a set of candidates in $[0, 0.8]$ (see Figure 4). Based on these observations, we set α at 0.16 empirically for training our model.

5 Evaluation

We evaluate all of our methods against several baselines on diverse data sets. We consider the time period of $1850 - 2005$ for the purpose of dating documents and evaluate our models on multiple data-sets described in Section 2. Since the tasks should get easier on long documents, we measure the performance of our models as a function of the length, Since the NYTIMES dataset only consists of the first paragraph of articles (about 100 tokens on average) we use the entire paragraph for evaluation on this dataset. Tables 1 and 2 show the Mean Absolute Error (MAE) over the NYTIMES and COHA datasets, from which we make the following observations:

- **Neologism Methods need relatively large documents to perform competitively**: The baseline NEO generally performs very poorly on short documents (of length 100 tokens). For example, on the COHA-FICTION dataset using 100 tokens, the MAE is 66.99 compared to BOOKPROP which yields an MAE of 44.57. On short documents NEO is easily misled due to lack of effective sample size. In contrast, observe that as the length of the document increases NEO's error reduces

#(Tokens)	#(Features)	MAE
BookProp	-	43.46
NEO	$\leq$ **200**	58.77
NEOProbMean	$\leq$ **200**	27.55
NEOProbMedian	$\leq$ **200**	28.22
NEOProbBayes	$\leq$ **200**	67.14
NaiveBayes	$\sim 200K$	23.69
NeuralDate (w/o reg.)	1000	22.80
NeuralDate	1000	**20.54**

Table 1: Mean Absolute Error on New York Times data. Note that neologism based methods whose feature set size is much smaller, performs competitively with NaiveBayes with a feature space of dimension of $200K$.

significantly (note Table 2 that for 2000 word documents on COHA-Fiction the mean absolute error is now 24.80).
Finally, the probabilistic models we propose extending NEO also perform better than Neo especially on short documents (for example, on COHA-Fiction for documents with 100 tokens the MAE for NeoProbMean is 32.40 compared to 66.99 for Neo). Similarly NeoProbMean and NeoProbMedian outperform NeoProbBayes on documents of upto 1000 tokens but NeoProbBayes almost always outperforms all of these on documents of length 2000, suggesting that NeoProbBayes needs a larger sample size to make effective predictions.

- **Deeper linguistic features boost performance**: We finally observe that including linguistic features like the words used in a simple Naive Bayes classifier consistently outperforms methods relying solely on neologisms. Further observe that the NeuralDate with the auto-correlation regularizer demonstrates superior performance over NeuralDate without regularization. Also note that our proposed neural based model NeuralDate also performs competitively and sometimes out-performs Naive Bayes (NB).

6 Analysis of Authors' Temporal Patterns

In this section, we conduct a preliminary analysis of temporal patterns associated with authors. Analyzing the linguistic styles of authors over time can improve authorship detection (Azarbonyad et al., 2015; McCarthy et al., 2006; Hansen et al., 2014) and also yield insights into the aging of authors (Lancashire and Hirst, 2009). In a similar vein, we apply our model to analyze patterns of rough linguistic styles of authors over time as we elaborate below.

We define the linguistic date LY of a book as the date predicted by our model based on the language used in the book and AY as the actual date the book was published. By measuring the correlation between the values of LY and AY for a predicted authors books, we obtain insight into the evolution of author's language usage over time.

We formulate and test two hypotheses:

(a) On an average, the second half of a book is written at a later date than the first half. To test this hypothesis, we considered 1252 books from Project Gutenberg and computed the fraction of instances where the second half of the book is assigned a date greater than the first half. We noted this fraction at $0.484 < 0.5$, but is not significant at the 0.05 level under a NULL model (where the expected fraction is 0.5), which is also evident from Figure 6. In summary, we cannot support the first hypothesis from our experiments.

(b) There is a positive correlation between the linguistic date and the actual publication date for authors. To test this hypothesis, we considered a set of six authors and obtained several of their books along with their publication dates from Project Gutenberg to conduct this author level analysis[4].

Figure 5 shows the correlations obtained between the linguistic dates and the actual dates for the books written by these 6 authors. First note that while our model over-estimates the linguistic date (LY) of a book by a few years, we observe a positive correlation between linguistic date (LY) and the actual date (AY). Second observe that authors like Arthur Conan Doyle, H. G. Wells and Mark Twain display a larger variation in their language over time than an author like P. G. Wodehouse. Further, note that all of the novels in the "Tom Sawyer" series by Mark Twain have very close linguistic dates, suggesting that these books have similar themes and literary styles as opposed to other books by Mark Twain.

These preliminary observations suggest that our methods can capture evolving linguistic pat-

[4] See Table 4 in in supplementary material for the detailed list.

Dataset	#(Tokens)	100	500	1000	2000
COHA-Fiction	BookProp	44.57	44.54	43.95	44.21
	NEO	66.99	34.74	27.39	24.80
	NEOProbMean	32.40	30.76	31.45	31.13
	NEOProbMedian	36.90	32.96	32.03	31.73
	NEOProbBayes	78.99	41.90	33.77	27.92
	NaiveBayes	**26.61**	**23.98**	**22.62**	**21.93**
	NeuralDate (w/o reg.)	37.56	30.71	28.96	27.97
	NeuralDate	35.66	30.02	27.81	26.96
COHA-NonFiction	BookProp	45.19	45.07	45.04	45.51
	NEO	57.99	30.75	24.84	22.86
	NEOProbMean	31.13	26.90	26.02	25.39
	NEOProbMedian	30.68	26.60	25.73	25.13
	NEOProbBayes	56.58	30.73	25.46	21.58
	NaiveBayes	**24.28**	**19.83**	**18.36**	**18.25**
	NeuralDate (w/o reg.)	27.91	23.57	22.29	21.60
	NeuralDate	25.21	20.07	20.38	20.09
COHA-News	BookProp	44.97	45.34	44.99	45.02
	NEO	39.86	20.39	19.80	20.26
	NEOProbMean	24.36	23.40	23.30	23.31
	NEOProbMedian	25.22	22.88	22.45	22.39
	NEOProbBayes	48.30	22.79	20.97	20.82
	NaiveBayes	21.35	17.21	16.64	16.60
	NeuralDate (w/o reg.)	20.40	16.04	15.43	15.34
	NeuralDate	**19.30**	**15.33**	**14.72**	**14.59**

Table 2: Mean Absolute Error of different models on COHA datasets as a function of number of tokens used for evaluation in each document. Note that our proposed neologism based methods that use a much smaller feature set perform competitively with NaiveBayes for long documents ($>=$ 500 tokens).

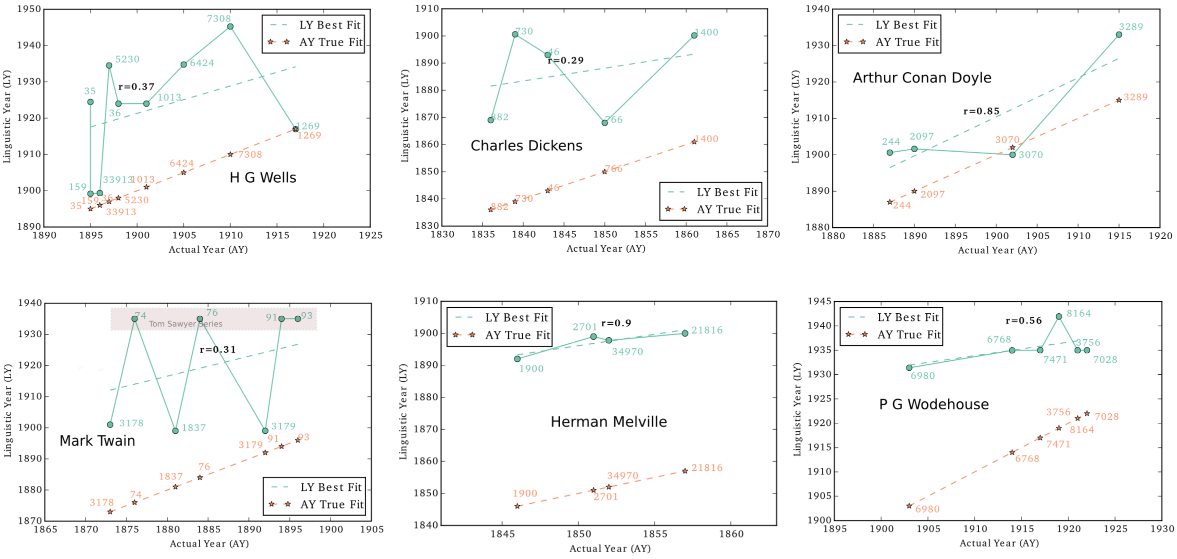


Figure 5: Temporal linguistic patterns of authors as revealed by an analysis of their books over time. Note that we observe a positive correlation (shown by the green line, not statistically significant at 0.05 level due to the small number of books per author) between the linguistic date and the actual date for the authors over time. The red line shows the trend when the linguistic dates and the actual dates are equal (for comparison). Note the positive correlation between linguistic date (LY) and the actual date (AY) for most of the authors we considered, suggesting that authors tend to evolve their linguistic style or exposition over time.The numbers are the ID's of the books on Project Gutenberg , the details of which are shown Table 4 in the supplementary material.

terns/styles of authors over time, One limitation of our analysis is that it is based on a relatively small set of authors and books written by each author since we were restricted to texts only available in the public domain. A large-scale analysis of several thousand authors and books and accounting for topic variation in authors would help strengthen/confirm our preliminary results.

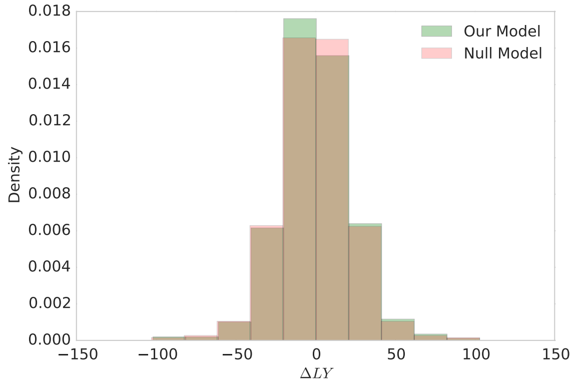


Figure 6: Distribution of ΔLY for the experiment testing Hypothesis (a). We observed no significant differences in the distributions obtained via the NULL model and our model (as seen by overlap in the figure), suggesting that we cannot provide evidence supporting Hypothesis (a).

7 Related Work

A large body of related work on the task of automatically dating texts exists in the field of temporal information retrieval. (Jong et al., 2005) proposed using temporal language models based on unigrams to date texts from the time period 1999 – 2005 on a dataset of Dutch newspaper articles. Several works then proposed incorporating additional features like lexical features, part-of-speech tagging, extraction of concepts and word sense disambiguation as well as using external knowledge bases (Kanhabua and Nørvåg, 2009; Garcia-Fernandez et al., 2011; Niculae et al., 2014; Zampieri et al., 2015, 2016). Our work is most closely related to the work of (Garcia-Fernandez et al., 2011) and (Zampieri et al., 2015). (Garcia-Fernandez et al., 2011) develop a model to date documents using both chronological methods with external knowledge and classification methods like using an SVM to date documents on a French Newspaper corpus while (Zampieri et al., 2015) propose a ranking based approach to temporal text classification. Our work differs from these in two aspects: (a) Both of these methods learn models on the respective domains they are evaluated on. In contrast, our proposed method seeks to learn a global model that can be applied across multiple domains without further tuning. (b) These methods uses neologisms to obtain a scoring function to date texts. In this work, we propose new probabilistic models to date texts by analyzing statistical patterns of introduction of new words over time. We also propose a novel penalty function to leverage temporal label structure, in order to learn neural models that account for this temporal structure in predictions. While general regularizers such as dropout, or label regularizers (Srivastava et al., 2014; Szegedy et al., 2016) can help prevent over-fitting, these methods do not encourage the model to make predictions that leverage this inherent structure.

Finally we also outline potential applications of our work to analyzing evolution of an author's language over time.

8 Conclusion

In this paper, we investigated the task of dating books on a large fine grained time scale (spanning 150 years) through the lens of neologisms introduced over time. We propose probabilistic models that effectively analyze the usage of neologisms, as well as a neural model to date documents across diverse domains. We demonstrate that these methods perform competitively with models that use deeper linguistic cues (which could use a feature space of more than thousands of features). Furthermore, our models are learned using only the Google Book NGrams, do not need any further tuning when evaluated on documents belonging to other domains and potentially enable researchers to obtain literary insights into language of authors over time.

References

Hosein Azarbonyad, Mostafa Dehghani, Maarten Marx, and Jaap Kamps. 2015. Time-aware authorship attribution for short text streams. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pages 727–730.

Colin Chase. 1997. *The dating of Beowulf*. 6. University of Toronto Press.

Mark Davies. 2002. *The Corpus of Historical American English (COHA): 400 million words, 1810-2009.*

Miles Efron. 2013. Query representation for cross-temporal information retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 383–392.

Martin Ehrensvärd. 1997. Once again: The problem of dating biblical hebrew∗. *Scandinavian Journal of the Old Testament* 11(1):29–40.

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

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*. volume 9, pages 249–256.

Niels Dalum Hansen, Christina Lioma, Birger Larsen, and Stephen Alstrup. 2014. Temporal context for authorship attribution. In *Information Retrieval Facility Conference*. Springer, pages 22–40.

Avi Hurvitz. 2000. Can biblical texts be dated linguistically? chronological perspectives in the historical study of biblical hebrew. *VETUS TESTAMENTUM-SUPPLEMENTS-* 80:143–160.

Adam Jatowt and Katsumi Tanaka. 2012. Large scale analysis of changes in english vocabulary over recent time. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, pages 2523–2526.

de FMG Jong, Henning Rode, and Djoerd Hiemstra. 2005. Temporal language models for the disclosure of historical text. Royal Netherlands Academy of Arts and Sciences.

Nattiya Kanhabua and Kjetil Nørvåg. 2009. Using temporal language models for document dating. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pages 738–741.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Ian Lancashire and Graeme Hirst. 2009. Vocabulary changes in agatha christie's mysteries as an indication of dementia: a case study. In *19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice*. pages 8–10.

Philip M McCarthy, Gwyneth A Lewis, David F Dufty, and Danielle S McNamara. 2006. Analyzing writing styles with coh-metrix. In *FLAIRS Conference*. pages 764–769.

Jean-Baptiste Michel et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–182.

Vlad Niculae, Marcos Zampieri, Liviu P Dinu, and Alina Maria Ciobanu. 2014. Temporal text ranking and automatic dating of texts. In *EACL*. pages 17–21.

Liudmila Ostroumova Prokhorenkova, Petr Prokhorenkov, Egor Samosvat, and Pavel Serdyukov. 2016. Publication date prediction through reverse engineering of the web. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, pages 123–132.

Mark F Rooker. 1996. Dating isaiah 40-66: What does the linguistic evidence say? .

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 2818–2826.

Ian Young and Robert Rezetko. 2016. *Linguistic dating of biblical texts*, volume 1. Routledge.

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). Denver, CO, USA*. pages 851–855.

Marcos Zampieri, Shervin Malmasi, and Mark Dras. 2016. Modeling language change in historical corpora: the case of portuguese. *arXiv preprint arXiv:1610.00030* .

A Supplemental Material

We present estimated year of popular usage (discussed in Section 3) for a small set of words in Table 3. We also show books processed for Project Gutenberg (discussed in Section 6) in Table 4.

Word	MR	FU
HIV	1987	1986
Hitler	1933	1934
LSD	1955	1950
Obama	2007	2006
SARS	2003	2003
Sputnik	1958	1957
electron	1905	1891
radio	1904	1907
television	1931	1907
transistor	1950	1948
walkman	1993	1979

Table 3: Example cases of estimated year of popular usage (MR) and actual year of first use [5] for different words from Google Book NGrams data. Note that in the majority of these words, our estimated year is close to the year of first usage and shows a small lag from the year of first use as expected.

No	Author	ID	Title	Year
0	Arthur Conan Doyle	244	A Study in Scarlet	1887
1	Arthur Conan Doyle	2097	The Sign of Four	1890
2	Arthur Conan Doyle	3070	The Hound of Baskervilles	1902
3	Arthur Conan Doyle	3289	The Valley of Fear	1915
4	H G Wells	35	The Time Machine	1895
5	H G Wells	33913	The Wonderful Visit	1895
6	H G Wells	159	The Island of Dr Moreau	1896
7	H G Wells	5230	The Invisible Man	1897
8	H G Wells	36	The War of the Worlds	1898
9	H G Wells	1013	The First Men in the Moon	1901
10	H G Wells	6424	A Modern Utopia	1905
11	H G Wells	7308	The History of Mr. Polly	1910
12	H G Wells	1269	The Soul of a Bishop	1917
13	Charles Dickens	882	Sketches by Boz	1836
14	Charles Dickens	730	Oliver Twist	1839
15	Charles Dickens	46	A Christmas Carol	1843
16	Charles Dickens	766	David CopperField	1850
17	Charles Dickens	1400	Great Expectations	1861
18	Mark Twain	3178	The Gilded Age	1873
19	Mark Twain	74	The Adventures of Tom Sawyer	1876
20	Mark Twain	1837	The Prince and the Pauper	1881
21	Mark Twain	76	The Adventures of Huckleberry Finn	1884
22	Mark Twain	3179	The American Claimant	1892
23	Mark Twain	91	Tom Sawyer Abroad	1894
24	Mark Twain	93	Tom Sawyer Detective	1896
25	Herman Melville	1900	Typee	1846
26	Herman Melville	2701	Moby Dick	1851
27	Herman Melville	34970	Pierre or the Ambiguities	1852
28	Herman Melville	21816	The Confidence Man	1857
29	P G Wodehouse	6980	Tales of St.Austin	1903
30	P G Wodehouse	6768	The Man Upstairs	1914
31	P G Wodehouse	7471	The Man with Two left feet	1917
32	P G Wodehouse	8164	My Man Jeeves	1919
33	P G Wodehouse	3756	Indiscretions of Archie	1921
34	P G Wodehouse	7028	The Clicking of Cuthbert	1922

Table 4: Table of Authors and their books analyzed over time.