Static vs. Contextual: Comparing Diachronic Word Embeddings for Medieval Latin Charters During the Norman Conquest

Yifan Liu Feixuan Chen Xinxiang Gao Gangquan Zhang
University of Toronto
{yifanliu.liu, feixuan.chen}@mail.utoronto.ca
{xinxiang.gao, gangquan.zhang}@mail.utoronto.ca

Abstract

The Norman Conquest significantly transformed administrative and societal structures in England. Medieval Latin charters, preserved in the University of Toronto's DEEDS database, offer a unique lens through which to examine these transformations using computational linguistic methods. The deployment of such methods requires high-quality dense representations of words from those periods; however, the existing embeddings for this low-resource historical language are limited. Although computational linguistics has explored both static and contextual embeddings in historical language contexts, their effectiveness in representing Medieval Latin charters continues to be questioned, particularly due to the significant linguistic shifts from the Anglo-Saxon to Norman periods. This paper marks the first systematic effort to implement and compare static and contextual embedding models on Medieval Latin charters. Our findings reveal that while contemporary studies often favor more flexible contextual embeddings such as BERT for their broad applicability, static models—when properly initialized with pre-trained modern embeddings—can actually perform better both intrinsically and extrinsically in this low-resource historical context. Furthermore, our models successfully identify meaningful semantic changes in subsequent intra-concept similarity tasks, establishing them as reliable tools for uncovering deeper insights into the linguistic and cultural changes of the Norman Conquest period.

1 Introduction

The Norman Conquest of 1066 is a pivotal event in English history, marking the beginning of Norman rule in England. A key outcome of the conquest was the importation of new administrative, cultural and linguistic practices by the Normans as evidenced in the change in the pattern of charter diplomatics preserved in the University of Toronto DEEDS (Documents of Early England Data Set) database (Gervers et al., 2018). Investigating the evolution of language within these charters over time presents a fascinating yet complex challenge. This process, known as Lexical Semantic Change (LSC), offers insights into the cultural and societal transformations of the period.

In the domain of computational linguistics, the majority of LSC studies have hitherto relied on high-quality distributional representations of words — vectors that capture a word's lexical context. This is often referred to as diachronic word embeddings. Throughout the early 2010s, the static word embedding method, such as the Continuous Skip-gram model (Mikolov et al., 2013) and the subword model (Bojanowski et al., 2017), served as the state-of-the-art approach for constructing diachronic embeddings (Kim et al., 2014; Hamilton et al., 2016; Xu et al., 2019; Xu and Zhang, 2021). In recent years, contextual representations based on deep neural networks, particularly BERT (Devlin et al., 2018) i.e., Bidirectional Encoder Representations from Transformers, have provided the advantage of dynamic, context-sensitive embeddings and richer semantic insights than their static counterparts. This advancement has opened new avenues for constructing diachronic embeddings (Kurtyigit et al., 2021; Kutuzov et al., 2022), with some efforts specifically targeting scarce historical languages (Qiu and Xu, 2022; Manjavacas Arevalo and Fonteyn, 2021; Beck and Köllner, 2023).

However, several issues still impede the application of embedding methods to Medieval Latin corpora. Firstly, our Medieval Latin corpora combined from the Anglo-Saxon and Norman periods contain only about 20k charters and 4M tokens[1]—a quantity significantly smaller than those studied in research on scarce historical languages. Secondly, Medieval Latin during the Norman Conquest is characterized by its rich, expansive vo-

[1] Our project is one of the largest in North America to maintain a collection of Medieval Latin charters, as detailed in Section 3.

cabulary and extensive borrowing from Old English and other local dialects. Meanwhile, the the existing research on diachronic word embeddings for Medieval Latin is extremely limited with the only known study to our knowledge being (Mehler et al., 2020). These raise methodological concerns regarding the applicability and performance of different word embedding models in this context. This paper, therefore, seeks to explore the following research questions:

1. Can we construct reliable representations of Medieval Latin during the Norman Conquest period in light of existing methods? Which type of representation — static or contextual — is more effective?

2. Can these representations be utilized to detect lexical semantic changes from the Anglo-Saxon to the Norman period, and if so, what types of changes can be identified?

To address the first research question, this paper will train various embedding models on Medieval Latin corpora from the Anglo-Saxon and Norman periods. The dataset description, including the corpora and pertinent metadata, is discussed in Section 3, and the training methods are outlined in Section 4. Then, we will evaluate different models both intrinsically and extrinsically as detailed in Section 5. To address the second question, we conduct a lexical semantic change analysis on a predefined set of target words, which is discussed in Section 6.

2 Related Work

2.1 Static Word Embeddings

The **Continuous Skip-gram** model, proposed by Mikolov et al. (2013), efficiently encodes contextual information of words by predicting their surrounding context words. In a similar vein, the **Continuous Bag-of-Words** model predicts the target words based on their context words. The **Subword model**, introduced by Bojanowski et al. (2017), enhances these methods by learning and aggregating context vectors through subword tokenization. This approach significantly improves the handling of out-of-vocabulary words and provides richer semantic representations, which is particularly beneficial for languages with complex morphologies. In the field of Lexical Semantic Change (LSC), the first research to integrate these prediction-based word vectors was conducted by Kim et al. (2014). Subsequent work by Hamilton et al. (2016) provided empirical evidence that neural-based diachronic embedding methods surpass traditional occurrence matrix-based approaches. Later studies have further refined these methods by incorporating subword models to enhance diachronic representations (Xu et al., 2019; Xu and Zhang, 2021).

The initialization plays an important role in static diachronic word embeddings. Word embeddings generated from different time periods may exhibit stochastic properties, which poses challenge for direct comparisons between embedding spaces. To address this, one effective strategy involves shared initialization. Kim et al. (Kim et al., 2014) constructed diachronic word embeddings for each year from 1850 to 2009, initializing each year's embeddings with the vectors from the previous year. In addition to enabling comparability between models, initializing word embeddings with pre-trained vectors also serves as a crucial strategy to address data scarcity. This approach is grounded in the principles of transfer learning to leverage pre-trained embeddings from a large corpus and fine-tuning them on a scarce dataset. Montariol and Allauzen (2019) trained diachronic embeddings for scarce corpora by introducing two initialization approaches: internal initiation, which initializes word embeddings for each time slice using vectors trained on the entire dataset, and external initiation, which utilizes pre-trained word embeddings from extensive external datasets.

2.2 Contexual Embeddings

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model that could leverage the power of transfer learning to enhance performance in downstream NLP tasks (Devlin et al., 2018). It allows for unique interpretations of words based on the specific contexts (input sequence) they appear. Researchers such as Hu et al. (2019), Giulianelli (2019), and Martinc et al. (2019) pioneered incorporating these contextual word representations, particularly BERT and its variants, into LSC studies. In the study of Martinc et al. (2019), they fine-tune a pre-trained BERT model to adapt it to the Liverpool FC corpus. This enabled them to obtain unique contextual representations for each instance of word usage across different time-slice subcorpora, which were then aggregated to represent all instances of a word within a single time-slice subcorpus. Because BERT inherently relies on time-specific contexts, this eliminated the need for the diachronic fine-tuning that traditional word embeddings typically require. Following this methodology, many have achieved state-of-the-art results in LSC using contextual embeddings or applied these techniques to a variety of languages beyond English (Kanjirangat et al., 2020; Rodina et al., 2021; Montariol and Allauzen, 2021; Kurtyigit et al., 2021; Kutuzov et al., 2022).

However, most existing pre-trained contextual representations are primarily based on contemporary corpora, leading to a predominant focus on recent time slices in LSC studies. Limited attention has focused on low-resources historical corpora. The first notable exploration of the potential biases in utilizing contemporary-trained contextual embeddings for historical data was conducted by Qiu and Xu (2022). They introduced histBERT, which adapts BERT to the historical American English corpus (COHA) dating back to 1910. Their findings demonstrated that histBERT outperforms the original BERT model in detecting linguistic semantic changes due to its ability to capture historical nuances. Nevertheless, while fine-tuning BERT models on contemporary data can enhance their performance on LSC detection tasks, it may still skew the fine-tuned model toward contemporary language. One proposed solution to this issue is the pre-training of a BERT model from scratch using historical data. Manjavacas and Fonteyn (2021; 2022) demonstrate this with MacBERTh, a model that pre-trained from scratch in English corpora from 1450 to 1900 CE and outperforms contemporary BERT model adaptations in handling similar historical corpora. Subsequent research expanded this approach to the German language, tracing back to 750 CE, by developing a GHisBERT model (Beck and Köllner, 2023). This period represents a time when the language differs even more substantially due to data stemming from another historical language stage. This paper demonstrates the possibility of leveraging contextual representations to construct diachronic embeddings during periods of significant linguistic transition.

3 Data

The University of Toronto DEEDS (Documents of Early England Data Set) Research Project was founded in 1975 by Michael Gervers, professor of History of the University of Toronto, to create a database of information culled from medieval property exchange documents, which would be of interest to social and economic historians.

In collaboration with Professor Michael Gervers, we have acquired two invaluable Medieval Latin datasets from the Anglo-Saxon and Norman periods. The dataset contains over 18,000 charters from the Anglo-Saxon and Norman periods, which are legal documents that record rights and privileges to properties. Table 1 provides a summary of the corpus data.

Period	**Anglo-Saxon**	**Norman**
Time Span	589-1066	1060-1310
Charters	1432	16976
Words	0.5M	3.6M
Vocabularies	61k	112k
Words Per Charter	341 (±290)	210 (±167)
Word Length	5.741 (±2.96)	5.743 (±2.87)

Table 1: Overview of the Medieval Latin corpora

Each document is accompanied by a dated year and metadata relevant to the corpora. The metadata in our dataset include: `Grant`, `Religion`, and `Confirmation`. Table 2 provides a detailed description of these metadata categories.

Metadata Type	**Description**
Grant	Indicating a legal document that formally recognizes the rights bestowed upon a recipient.
Religion	Pertains to the monastery or religious order that issued the documents.
Confirmation	Indicating a document that reaffirms existing grants and agreements, ensuring their continuity over time.

Table 2: Description of the metadata

The datasets underwent an initial cleaning process during their digitalization in the 1990s, led by Michael Margolin under the guidance of Professor Gervers. We further standardized the spelling of characters and words based on expert recommendations. In our preprocessing pipeline, we attempt to standardize words that appear fewer than five times if a highly similar word (spelling similarity greater than 0.85) appears more than five times. We define spelling similarity between a pair of words based on the Levenshtein distance:

$$\text{similarity}(w, w') = 1 - \frac{\text{distance}(w, w')}{\text{len}(w) + \text{len}(w')}$$

where *distance*(w, w') represents the Levenshtein distance between the words w and w', and *len*(w) and *len*(w') are the lengths of the words w and w', respectively.

4 Methods

4.1 Static Word Embeddings

We utilized the Continuous Skip-gram model with subword information to generate static word embeddings for the Anglo-Saxon and Norman periods (Mikolov et al., 2013; Bojanowski et al., 2017). Following the approach proposed by Montariol and Allauzen (2019), we adopted both internal and external initialization methods. For internal initialization, denoted `fasttext-internal`, we trained a joint model on the combined corpora from both periods, conducting the training over 30 epochs[2] with a learning rate decreasing from 0.025 to 0.0001 and using a 5-gram setting (the average word length in the corpus of 6.5 characters). Each period was then trained individually for an additional 30 epochs with the same hyperparameters. For external initialization, labeled `fasttext-external`, we utilized pre-trained modern Latin[3] word embeddings from Grave et al. (Grave et al., 2019), which were trained using CBOW with position-weights in dimension 300, character n-grams of length 5, a window size of 5, and 10 negatives. Although word morphology and semantic changes occur from Medieval Latin to Modern Latin, the subword model leverages the information stored in n-grams (e.g., prefixes, stems), which may remain unchanged since their morpheme inception. We then fine-tuned the embeddings on Anglo-Saxon and Norman corpus separately for another 30 epochs using the same hyperparameter settings with Grave et al. Both the internal and external embeddings were trained using vector sizes of 100 and 300. All models were implemented using the FastText module in the Gensim library (Řehůřek and Sojka, 2010).

4.2 Contexual Embeddings

In alignment with the methodologies by Manjavacas and Fonteyn (2021) and Beck and Köllner (2023) for historical English and German texts, we introduce two distinct BERT-based contextual embedding variants for comparison. The `pretrained-bert` model was developed through pre-training from scratch on the medieval Latin corpora (i.e., both Anglo-Saxon and Norman periods corpus) described Section 3. We adopted Beck and Köllner's (2023) training hyperparameter configuration given both of our limited corpus size. Specifically, the model was trained with 12 hidden layers, each with a dimensionality of 768, alongside 12 attention heads and a 32,000 token vocabulary.[4]The model was pre-trained on a masked language modelling (MLM) task by randomly masking 15% of the corpus. Training proceeded over 10 epochs, with batches of 8 and the implementation of gradient accumulation techniques. The model took 3 hours to train on a 16GB NVIDIA Tesla V100 GPU.

For comparative analysis, we fine-tuned an existing BERT model pre-trained on a modern Latin corpus, denoted `finetuned-bert`. The selected pre-trained model, LuisAVasquez/simple-latin-bert-uncased[5], was originally trained on corpora from the Classical Language Toolkit (CLTK) with the MLM task. This model features 12 attention heads, 12 hidden layers and a hidden size of 768, consistent with the standard BERT architecture. We continually trained this model from its last checkpoint with our medieval Latin corpus over an additional 4 epochs. Throughout the fine-tuning process, we adhered to the original model's hyperparameter configuration to ensure consistency. The model took 2 hours to train on a 16GB NVIDIA Tesla V100 GPU.

For both models, we pre-train our tokenizer before feeding the historical data. This accounts for the complex and archaic linguistic patterns and

[2] Various training epochs were tested; however, modifications did not significantly affect the model's performance, thus are not detailed further due to space constraints.

[3] We are aware of some pre-trained Medieval Latin word vectors (Mehler et al., 2020); however, they are not formatted to support continued training.

[4] We also implemented two alternative configurations: a smaller model with a hidden size of 256, 4 attention heads, and 4 hidden layers, and a model incorporating an additional Medieval Latin corpus from France. Neither modification significantly affected the model's performance; thus, they are not reported due to page constraints.

[5] We could also try Latin BERT (Bamman and Burns, 2020) or multi-lingual BERT (Devlin et al., 2018); however, we were constrained by limited computational resources.

diverse word forms. The tokenizer was trained with the same hyperparameter settings as those outlined by Beck and Köllner's (2023). Specifically, we utilize HuggingFace's BertWordPieceTokenizer with 32000 vocabularies and a max sequence length of 512.

Some subsequent sections involve distilling static embeddings from BERT-based models, following the methodology initially described by Martinc et al. (Martinc et al., 2019). We begin by extracting the average of the last four encoder layers of the BERT model for a sequence s containing n tokens, denoted as w_s. Next, we split w_s into n segments to obtain contextual embeddings for each token occurrence. These token embeddings are then grouped by their corresponding words, as identified by overlapping offset mappings in the tokenization process. We average the embeddings within each group to derive word-level embeddings for each occurrence of the words across the corpus. Finally, these embeddings are averaged to generate a dense representation for each word.

Additionally, we create two more versions of the model by reducing the vector dimension to 100 and 300, respectively, using the PCA implementation from Scikit-learn (Pedregosa et al., 2011).

5 Evaluation

Word embedding assessments generally fall into two categories: intrinsic and extrinsic evaluations. Intrinsic evaluations examine if embeddings offer semantically coherent representations for corpus words, typically through word similarity and analogy tests. On the other hand, extrinsic evaluations verify the adaptability of embeddings to diverse downstream tasks. This paper aims to provide a systematic comparison of different embedding models' on the medieval Latin corpus, so both intrinsic and extrinsic evaluation will be conducted.

5.1 Metadata Classification

The dataset comes with high-quality metadata labels[6] that reflect the political and cultural context of the charters. These labels serve as an apt basis for downstream text classification tasks, which could be used to assess the efficacy of different embedding models in generating meaningful representations for the diverse aspects of the data. The metadata includes both binary (`Grant`, `Confirmation`) and multi-class labels (`Religion`), the meanings of which are already discussed in Section 3. Table 3 provides summary statistics for these metadata labels.

Category	**Grant**	**Confirmation**	**Religion**
Obs.	16976	16976	7489
Type	Binary	Binary	Multi
Distribution	1: 20.95%	1: 12.79%	Benedictine: 38.27%
	0: 79.05%	0: 87.21%	Augustinian: 38.68%
			Cluniac: 3.02%
			Cistercian: 10.23%
			Hospitaller: 3.3%
			Omitted Others: 6.51%

Table 3: Summary of the Norman metadata

5.1.1 Evaluation Methods

Previous studies, such as Thawani et al. (2019) and Wang et al. (2020), have established the methodological framework for the extrinsic evaluations on various classic word embeddings like Word2Vec, GloVe, and FastText and contextual embeddings like BERT and ELMo in downstream text classification and inference tasks.

Building on these methodologies, our evaluation task is designed as follows. We first tokenize an input charter sequence s into a sequence of words/tokens. The embedding layer e processes these tokens by utilizing each word embedding to be evaluated as fixed feature extractors and generates an embedding matrix $\mathbf{o}$ for the sequence. Specifically, static embeddings assign vector values to each token to form the matrix representations; contextual embeddings from BERT extract the second-last layer outputs as the sequence's embedding matrix. This representation then feeds into a classifier. Our model employs a BiLSTM architecture for the classifier with a hidden layer h (256 hidden units, dropout rate: 0.3) and a fully connected layer f, following the standard implementation in PyTorch's Neural Networks module (Paszke et al., 2019). The output logits $\mathbf{g}$ are transformed into predicted probabilities through a sigmoid function. For model optimization, we employ the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001, over 20 training epochs, and a batch size of 128.

To enhance the robustness of the score we utilized a stratified 5-fold cross-validation. We report the mean accuracy and F1 score across all 5-folds for each text classification task.

[6] Metadata labels are available only for the Norman period; therefore, the results discussed in this section pertain solely to that period.

5.1.2 Evaluation Results

Table 4 reports the evaluation results. Based on these results, we can draw the following conclusions:

Static word embeddings: The external-fasttext model significantly outperforms the internal-fasttext model across all classification tasks. This might be attributed to the high-quality pre-trained modern Latin embeddings, which adapt well to the Medieval Latin corpus due to their shared morphological components. The impact of embedding size varies between models. For the internal-fasttext models, an embedding size of 300 leads to better performance than the 100 counterpart, likely because the larger size captures more semantic information. For the external-fasttext models, the effect of a larger embedding size is slightly negative but not significant, possibly because the pre-trained embeddings with a size of 100 already encode abundant semantic information for the metadata prediction tasks.

Contextual embeddings: The pretrained-bert performs significantly better than the finetuned-bert. This aligns with findings by Manjavacas and Fonteyn (2021) and Beck and Köllner (2023), but contrasts with the performance of static models where the fine-tuned versions on modern Latin perform better. There are two possible explanations for these results. Firstly, pre-training BERT from scratch on historical language could avoid biases towards contemporary language. Unlike fine-tuning traditional static embedding models, fine-tuned historical BERT models are more likely biased towards contemporary language, as the BERT model retains much of its initial training on modern data. Another possibility is that the training quality of the modern Latin-BERT model used for fine-tuning was suboptimal. We considered experimenting with different base BERT models that support Latin for a more thorough comparison, such as the modern Latin-BERT by Bamman and Burns (2020) or the multi-lingual BERT by Devlin et al. (2018). However, these experiments were not conducted in this study due to resource constraints.

Static vs. Contextual: The external-fasttext outperforms all BERT models in binary classification tasks (i.e., Grant, Confirmation). For the multi-class classification task (i.e., Religion), the pretrained-bert achieves a significantly higher F1 score than the static models, though with lower prediction accuracy. This discrepancy might be explained by the nature of the metadata: Confirmation and Grant could be identified based on key sentences within the charters, while Religion metadata may involve more complex information. Static embeddings are inherently suited to tasks that involve recognizing word occurrences. This was further evidenced by evaluating the distill-pretrained-bert model[7], a static version of BERT embeddings. It performed better on Confirmation and Grant than the pretrained-bert (though still inferior to the fasttext-external) and performed worse on Religion. Given the relative performance of all models, we conclude that the fasttext-external offers the most adaptable representations for Medieval Latin corpora.

5.2 Concept Similarity

5.2.1 Evaluation Methods

There are no existing labelled word similarity tests for Medieval Latin to automatically evaluate the model, which are typically included in previous works (Levy and Goldberg, 2014; Bojanowski et al., 2017). We instead assesses the semantic relationships captured by the models as sources of evidence of the intrinsic quality of word embeddings. Specifically, we measure the inter-concept similarity at each time stage, which quantifies how closely the average embedding of each concept aligns with the average embeddings of all other concepts.In this section, we evaluate BERT model only through the distillation method outlined in Section 4 to calculates the concept similarity. Specifically, to ensure a fair comparison between the BERT and static models, we employ distilled BERT embeddings with hidden dimensions of 100 and 300 (i.e., matching the dimensions used in the static models) as detailed in Section 4.

Prof. Gervers helped identify seven significant religious and royal concepts present in both

[7] We limit our experiments to distilling the BERT model from the pre-trained version with the default hidden size of 768. While it is possible to apply the same methodology to fine-tuned BERT models and to versions with reduced vector dimensions of 100 and 300, we have not done so due to constraints on time and resources.

Model	**Grant**		**Religion**		**Confirmation**	
	Accuracy	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
external-fasttext-100	**88.84 ± 0.49**	**72.06 ± 2.30**	**56.15 ± 5.80**	18.72 ± 4.65	91.16 ± 0.77	**65.69 ± 2.67**
external-fasttext-300	88.39 ± 0.32	70.51 ± 2.23	55.83 ± 9.89	17.59 ± 6.59	**91.63 ± 0.27**	65.28 ± 2.23
internal-fasttext-100	80.88 ± 2.65	46.52 ± 15.16	50.90 ± 2.82	15.08 ± 1.60	88.27 ± 1.89	40.08 ± 18.76
internal-fasttext-300	82.17 ± 0.91	54.90 ± 4.57	51.26 ± 2.02	15.89 ± 1.51	88.28 ± 0.99	41.48 ± 15.06
pretrained-bert	85.92 ± 0.22	63.39 ± 2.72	51.21 ± 0.89	**28.60 ± 2.21**	88.55 ± 0.99	45.18 ± 8.47
finetuned-bert	84.20 ± 0.38	58.41 ± 4.25	45.75 ± 0.93	16.52 ± 1.04	87.70 ± 0.83	33.68 ± 6.51
distill-pretrained-bert	86.61 ± 0.40	67.23 ± 1.38	53.18 ± 6.11	15.68 ± 2.53	89.96 ± 0.34	56.79 ± 2.18

Table 4: Evaluation results: accuracy and F1 for metadata classification in Norman period

Anglo-Saxon and Norman corpora. Table 5 details these concepts. For each time stage, we computed the cosine similarity between each pair of concepts. Prof. Gervers then conducted a qualitative evaluation of the trends observed across the models.

Concept	**Anglo-Saxon** **n**	**Norman** **n**
FACTA ‘act’	104	4187
REGIS ‘king’	1129	13112
DOMINI ‘lord’	816	14232
PERPETUAM ‘perpetual’	146	3089
ELEMOSINAM ‘charity’	55	2997
EPISCOPUS ‘bishop’	3665	4197
DIE ‘day’	506	8686
All	6261	48673

Table 5: Target concepts at each language stage

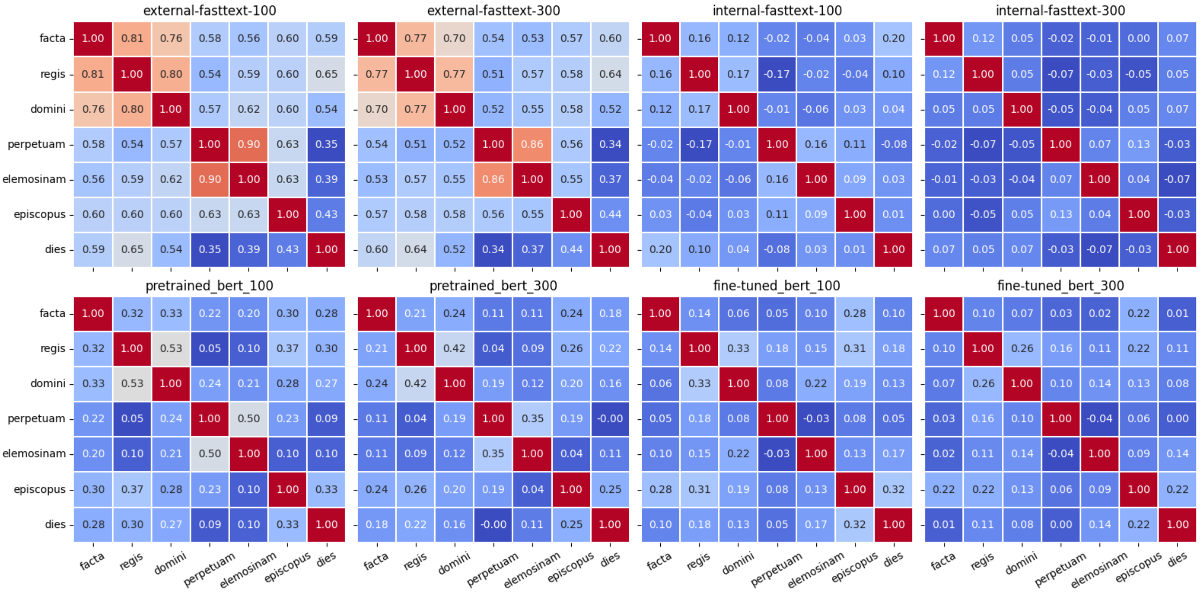
5.2.2 Evaluation Results

Figures 1a and 1b illustrate the inter-concept similarity for all targeted concepts across different language models and historical periods. The following conclusions can be drawn from the results:

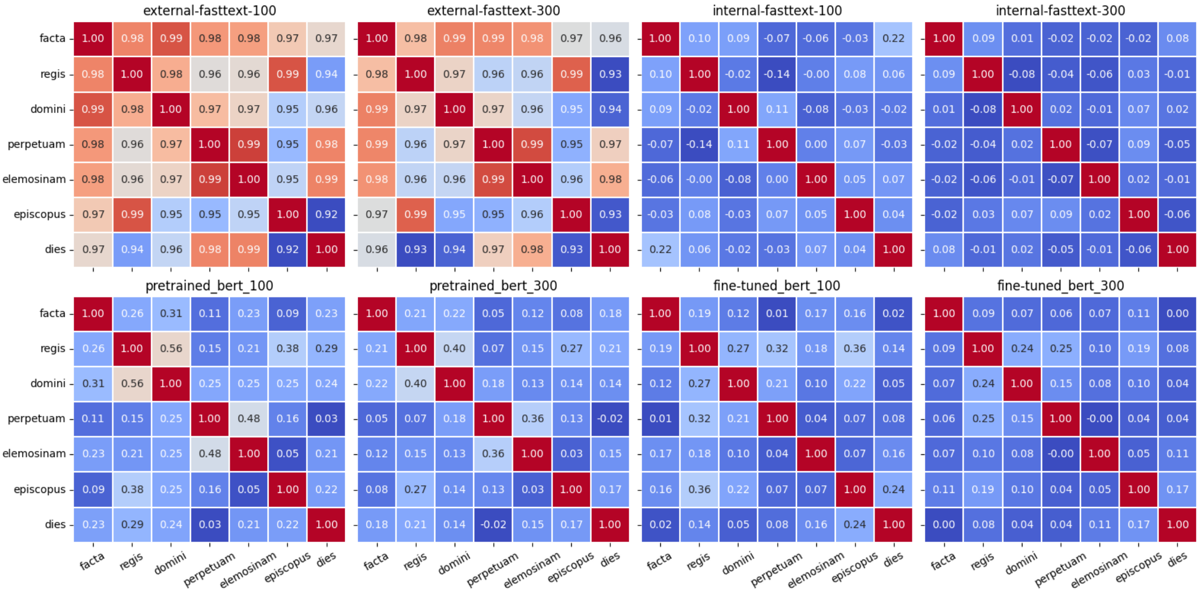
Norman Period The `external-fasttext` model demonstrates the strongest capability for capturing semantic relationships during the Norman period. It provides the most reasonable range (0.35 to 0.81) and correlation of concept similarity scores, which reflects the model most effectively captures the varying degrees of relatedness among concepts. For example, it identifies a high similarity between REGIS and DOMINI, which both represent authoritative figures. It also correctly associates PERPETUAM with ELEMOSINAM, which frequently co-occurring in grant documents to denote long-lasting grant bestowed by royalty. It also appropriately distinguishes the low similarity between DIE and other concepts, as it is neither religious nor royal. However, the model's tendency to draw strong associations between DIE and REGIS, as well as between FACTA and REGIS / DOMINI, is difficult to justify. The `pretrained-bert` models exhibit similar trends but with a slightly narrower similarity range (from 0.05 to 0.53). Notably, they avoid some of the unreliable relationships identified by the `external-fasttext` models. The `internal-fasttext` model, while capturing comparable relationships, offers a similarity range that is too low to be robust (from -0.17 to 0.2). The `finetuned-bert` model not only displays a limited range of similarity but also presents unusual relationships, such as the exceptionally low similarity between PERPETUAM and ELEMOSINAM. The potential reasons for this include the suboptimal training of the base Latin-BERT or its strong modern Latin bias. This can be further investigated by experimenting with other base model setups. All models with a hidden size of 100 tend to find higher inter-concept similarity compared to those with a hidden size of 300, likely due to the curse of dimensionality.

Anglo-Saxon All models show similar behaviors in the Anglo-Saxon period, except for `external-fasttext`. The latter shows rather high similarity between each pair of words. This suggests a failure to discern meaningful information from the Anglo-Saxon data, possibly because Medieval Latin underwent substantial changes due to the Norman conquest. Therefore, Medieval Latin from the Norman period more closely resembles modern Latin than its Anglo-Saxon counterpart, making modern Latin a potentially unsuitable initialization point for Anglo-Saxon embedding models. This does not detract from the potential of external initialization methods; indeed, a more suitable external initialization could be identified for the Anglo-Saxon corpus.[8]

[8] Due to time constraints, we were unable to identify a



(a) Inter-concept similarity for Norman period



(b) Inter-concept similarity for Anglo-Saxon period

6 Semantic Change Detection

Detecting lexical semantic change (LSC) without a human-labeled dataset is exceptionally challenging. As this represents the first research endeavor focused on LSC in Medieval Latin, no gold standard data exists for conducting a comparable ground truth evaluation. Therefore, we instead focus on a qualitative assessment of the validity of the target concepts. Similarly to the inter-concept similarity analysis, we utilize distilled BERT models to enable the calculation of concept similarity.[9]

Figure 2 illustrates the intra-concept similarity for various target words across different linguistic stages. Different models tend to produce varying numerical values for intra-concept similarity. BERT-based models typically show the highest similarity values, as the Anglo-Saxon and Norman corpora are trained together in these models. The numerical values from the `external-fasttext` model may be considered unreliable since this model fails to provide robust embeddings for the Anglo-Saxon corpus.

Among all models, a consistent pattern emerges

better initialization. This will be addressed in future work.

[9] We do not include the dimensionality reduction version of the distilled BERT, as PCA rotation of the embedding space for the Anglo-Saxon and Norman models would render them incomparable.

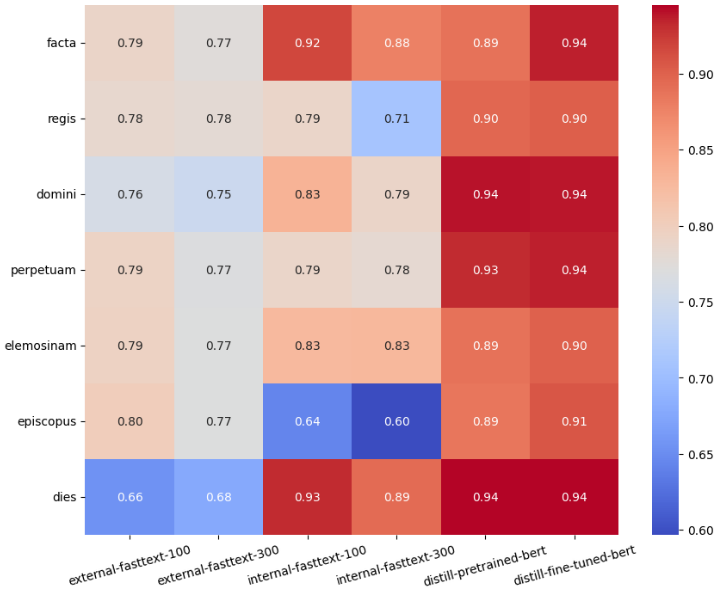


Figure 2: Intra-concept similarity calculated by each model across periods

where EPISCOPUS and REGIS are the concepts that have undergone the most significant semantic changes. This aligns with our historical understanding of the Norman Conquest, which marked a profound transformation in the governance and ecclesiastical structure of England. The conquest led to the centralization of power, replacing the relatively decentralized Anglo-Saxon governance with a feudal system centered around the monarchy and the Church. Additionally, the concepts DIE and DOMINI exhibit the least semantic change; ’day’ continues to represent time, a concept that remains constant across periods, while DOMINI, describing the lord or God, is a notion that remains relatively unchanged in religious contexts. Therefore, our models are effectively detecting meaningful semantic changes that mirror historical shifts in the unlabeled setting, though further validation with labeled datasets is advisable for future studies.

7 Conclusion

This paper represents the first research endeavor to systematically implement and compare static prediction-based word embeddings and contextual embeddings for Medieval Latin charters during the Norman conquest period. Through intrinsic and extrinsic evaluations, we demonstrate that static FastText models, initialized with pretrained modern Latin embeddings, are the most effective and adaptable for encoding semantic information in the Norman period. This shed a light on the ongoing debate about the efficacy of static versus contextual embeddings in low-resource historical data, supporting the notion that properly initialized static models can outperform more data-intensive language models in sparse historical corpora. Among the BERT-based models, we observed generally satisfactory performance, even under extremely low-resource conditions. Our results indicate that pre-training a BERT-based model from scratch with relevant historical data yields more adequate results than mere fine-tuning. In a subsequent semantic change detection analysis, we show that our model are able to extract meaningful semantic changes from the Anglo-Saxon to the Norman periods that reflect actual social and cultural transformations during those times.

Future Work

This research opens new avenues for historical linguistic to explore historical Medieval Latin charters. Future studies could leverage the model developed and use them in various downstream tasks in construct the Medieval Latin dataset. As illustrate in this paper, scholars could utilize the lexical semantic change detection framework discussed in this paper as a knowledge discovery process to understand the social and cultural change in that periods. They could also use this model to build automatic metadata labeller for the new charters found.

Future studies could build upon the models developed here for various downstream tasks in constructing Medieval Latin datasets. As demonstrated, scholars could utilize the lexical semantic change detection framework discussed to facilitate knowledge discovery processes for better understanding the social and cultural shifts of the era. These models could aid in the automatic labeling of metadata for newly discovered charters.

Despite its contributions, this study is not without limitations. We aim to explore various external initializations for the FastText model that align better with the vocabulary from the Anglo-Saxon period to further support our conclusion about the superiority of this model. Further, we plan to provide a more diverse set of baseline models for fine-tuned BERT-based methods to enable fair comparisons between pre-trained and fine-tuned approaches. As a pioneering investigation into embeddings in Medieval Latin charters, our study's absence of a gold standard dataset for intrinsic evaluations (e.g., word similarity tests, word analogy tests) and quantitative semantic change analysis is a significant limitation. In future work, we plan to collaborate further with Professor Michael Gervers and other scholars on Medieval Latin to construct a gold standard dataset. This will not only allow for quantitative evaluations of our findings but also serve as a benchmark for future research.

Acknowledgement

Each member was responsible for different aspects of the research. Yifan Liu developed the overall idea of the project, implemented the code for the BERT model, and wrote the introduction, related work, and BERT-related methods sections. Feixuan Chen handled the intrinsic evaluation of the models and conducted the lexical semantic change detection, including both coding and writing the relevant sections of the paper. Xinxiang Gao focused on the static word embedding models, conducted a literature review on these methods, implemented the training procedures, and wrote the corresponding sections of the paper. Gangquan Zhang was responsible for the extrinsic evaluation of the models, wrote the relevant sections, and developed the code to extract static BERT models. We also thank Prof. Michael Gervers and Dr. Gelila Tilahun for providing us with access to DEEDS project database, sharing their knowledge of Medieval Latin from the Anglo-Saxon and Norman periods, and manually evaluating the inter-concept similarity test results.

References

David Bamman and Patrick J Burns. 2020. Latin bert: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.

Christin Beck and Marisa Köllner. 2023. Ghisbert–training bert from scratch for lexical semantic investigations across historical german language stages. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 33–45.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Michael Gervers, Gelila Tilahun, Shima Khoshraftar, and Roderick A Mitchell. 2018. The dating of undated medieval charters. *ARCHIVES: The Journal of the British Records Association*, 53(137):1–33.

Mario Giulianelli. 2019. Lexical semantic change analysis with contextualised word representations. *Unpublished master's thesis, University of Amsterdam, Amsterdam*.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3899–3908.

Vani Kanjirangat, Sandra Mitrovic, Alessandro Antonucci, and Fabio Rinaldi. 2020. Sst-bert at semeval-2020 task 1: Semantic shift tracing by clustering in bert-based embedding spaces. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 214–221.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

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

Sinan Kurtyigit, Maike Park, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Lexical semantic change discovery. *arXiv preprint arXiv:2106.03111*.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2022. Contextualized language models for semantic change detection: lessons learned. *arXiv preprint arXiv:2209.00154*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. pre-training language models for historical languages. *Journal of Data Mining & Digital Humanities*, (Digital humanities in languages).

Enrique Manjavacas Arevalo and Lauren Fonteyn. 2021. MacBERTh: Development and evaluation of a historically pre-trained language model for English (1450-1950). In Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter, editors, *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 23–36, NIT Silchar, India, December. NLP Association of India (NLPAI).

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2019. Leveraging contextual embeddings for detecting diachronic semantic shift. *arXiv preprint arXiv:1912.01072*.

Alexander Mehler, Bernhard Jussen, Tim Geelhaar, Alexander Henlein, Giuseppe Abrami, Daniel Baumartz, Tolga Uslu, and Wahed Hemati. 2020. The frankfurt latin lexicon: From morphological expansion and word embeddings to semiographs. *arXiv preprint arXiv:2005.10790*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Syrielle Montariol and Alexandre Allauzen. 2019. Empirical study of diachronic word embeddings for scarce data. *arXiv preprint arXiv:1909.01863*.

Syrielle Montariol and Alexandre Allauzen. 2021. Measure and evaluation of semantic divergence across two languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1247–1258.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Wenjun Qiu and Yang Xu. 2022. Histbert: A pretrained language model for diachronic lexical semantic analysis. *arXiv preprint arXiv:2202.03612*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Julia Rodina, Yuliya Trofimova, Andrey Kutuzov, and Ekaterina Artemova. 2021. Elmo and bert in semantic change detection for russian. In *Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020, Revised Selected Papers 9*, pages 175–186. Springer.

Avijit Thawani, Biplav Srivastava, and Anil Singh. 2019. Swow-8500: Word association task for intrinsic evaluation of word embeddings. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 43–51.

Congcong Wang, Paul Nulty, and David Lillis. 2020. A comparative study on word embeddings in deep learning for text classification. In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, pages 37–46.

Yang Xu and Zheng-sheng Zhang. 2021. Historical changes in semantic weights of sub-word units. *Computational approaches to semantic change*, pages 169–187.

Yang Xu, Jiasheng Zhang, and David Reitter. 2019. Treat the word as a whole or look inside? subword embeddings model language change and typology. In Nina Tahmasebi, Lars Borin, Adam Jatowt, and Yang Xu, editors, *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 136–145, Florence, Italy, August. Association for Computational Linguistics.