Towards Detecting Lexical Change of Hate Speech in Historical Data

Sanne Hoeken[*1], **Sophie Spliethoff**[*2], **Silke Schwandt**[2], **Sina Zarrieß**[1] and **Özge Alaçam**[1,3]
[1]Computational Linguistics, Dept. of Linguistics, Bielefeld University
[2]Faculty for History, Philosophy and Theology, Bielefeld University
[3]Center for Information and Language Processing, LMU Munich
{sanne.hoeken, sophie_jasmin.spliethoff, silke.schwandt, sina.zarriess, oezge.alacam}@uni-bielefeld.de

Abstract

The investigation of lexical change has predominantly focused on generic language evolution, not suited for detecting shifts in a particular domain, such as hate speech. Our study introduces the task of identifying changes in lexical semantics related to hate speech within historical texts. We present an interdisciplinary approach that brings together NLP and History, yielding a pilot dataset comprising 16th century Early Modern English religious writings during the Protestant Reformation. We provide annotations for both semantic shifts and hatefulness on this data and, thereby, combine the tasks of Lexical Semantic Change Detection and Hate Speech Detection. Our framework and resulting dataset facilitate the evaluation of our applied methods, advancing the analysis of hate speech evolution.[1]

> "the **foxes** have hooles, the birdes of the aire have nestes, but the sonne of manne hathe not wherin to lay his heade."
> *Richard Tracy (1544)*

> "so these **foxes** conceive mischiefe and bring foorth most monstrous and cruel wickednesse; both by open violence and by secret treacherie."
> *Edwin Sandys (1585)*

Figure 1: Example of 'foxes', for which our study found that the use of its hateful meaning *increased* between the periods of 1530-1553 and 1580-1603.

1 Introduction

The present research landscape on lexical change in NLP predominantly focuses on generic language evolution, targeting shifts in meaning for a set of words that span a wide spectrum of vocabulary (Schlechtweg et al., 2020; Basile et al., 2020). This approach falls short of modeling meaning shifts in specific domains or dimensions of meaning (e.g. hatefulness), which is often of interest when applying language change detection in disciplines beyond linguistics, i.e. in social sciences and humanities. For instance, historians investigating religious conflicts between Protestants and Catholics during the English Reformation may be particularly interested in the dynamics of polemical expressions (Steckel, 2018; Schwerhoff, 2020), which exist within a limited subset of the lexicon. In this paper, we present a first step towards the detection of meaning shifts within a particular subdomain. Specifically, our focus is on the domain of hate, aiming to uncover, for instance, change in hateful usage of the term 'foxes' during the 16th century as illustrated in Figure 1.

Lexical Semantic Change Detection (LSCD) is currently the predominant approach to modeling meaning shift in NLP. LSCD methods are typically designed to observe shifts in word usage, targeting a word's denotative meaning within evaluation data encompassing general language sources such as newspapers and books (Schlechtweg et al., 2020; Zamora-Reina et al., 2022). Target words are selected from the full vocabulary, often guided by etymological and historical dictionaries. Following this, well-developed techniques detect semasiological (from term to concept) variation by determining to what extend a word has shifted its meanings *somehow*. While certain more interpretable methods could offer deeper insights into the nature of individual shifts, by e.g. looking into usage clusters or word substitutes (Montariol et al., 2021; Card, 2023), current LSCD approaches have not demonstrated the ability to detect shifts within specific semantic subdomains, such as hate, which could be considered as an onomasiological perspective (from concept to term).

The development of Hate Speech Detection (HSD) systems, on the other hand, does address

[*]These authors contributed equally to this work.
[1]The published dataset and code used can be found at https://github.com/SanneHoeken/DigHist

Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, pages 100–111
December 6, 2023

the identification of lexical items used to convey hateful meanings (e.g. Gitari et al., 2015; Bassignana et al., 2018; Davidson et al., 2017). However, the evolution of these expressions often remains unexplored (with McGillivray et al. (2022) being a rare exception). Although hate speech lexicons are frequently integrated into these systems, their application is most prevalent within limited temporal scopes, such as short-term social media data sets.

Our study introduces the task of detecting lexical semantic change of hate speech in historical texts. Such changes can involve an increase or decrease in hatefulness, or even the acquisition of an entirely new hateful sense. To address this task, we present an interdisciplinary framework, that brings together NLP and History. More specifically, we use and combine methods, annotation and evaluation procedures for Lexical Semantic Change Detection (LSCD) and Hate Speech Detection (HSD) in the context of historical data. The resulting dataset, consisting of 16th century Early Modern English religious writings in the context of the Protestant Reformation, is enriched with annotations of both lexical semantic changes and lexical hatefulness. In conclusion, our paper presents a 1) task, 2) dataset and 3) methodological framework facilitating the evaluation of computational approaches for identifying shifts in hateful word meanings.

2 Related Work

2.1 Historical text analysis

Semantic changes in historical polemical writing have not yet been targeted with the help of computational methods; instead, historians and literary researchers focused on qualitative approaches, such as close reading methods, in order to work out characteristics of polemical speech (Bevan Zlatar, 2011; Almasy, 2008). Moreover, Steckel (2018) and Schwerhoff (2020) provide first conceptualisations of historical polemics as a research instrument, and Dröse (2021) shows how interwoven these writings were with medial changes. Nevertheless, there have been approaches to apply NLP methods in the fields of Digital Humanities and Digital History already, which demonstrate that using digital methods to deal with historical texts does not only enable us to generate new findings and process larger amounts of textual data. As highlighted by Schwandt (2018), an interdisciplinary approach combining computational methods and practices with historical research also changes the way we perceive and interpret text and allows for new research questions.

2.2 Lexical Semantic Change Detection (LSCD)

In LSCD, a diverse range of methods has been employed, leveraging various language modeling techniques, including count-based models, static word embedding models, and contextualized language models. Tahmasebia et al. (2021) or Montanelli and Periti (2023) provides a comprehensive overview for further reading.

The evaluation of LSCD methods has been challenging due to the lack of large-scale annotated data. The first SemEval shared task on LSCD in 2020 provided one of the few available larger-scale human-annotated evaluation datasets (Schlechtweg et al., 2020). Interestingly, the results on this task demonstrated that methods utilizing static word embedding models, e.g. Hamilton et al. (2016), outperformed other approaches, including those using BERT-based models (Kutuzov and Giulianelli, 2020). More recently, several methods based on contextualized models have shown greater success, either by extracting representations from a Transformer-based model fine-tuned on Word Sense Disambiguation (Rachinskiy and Arefyev, 2022), or relying on the most probable substitutes for masked target terms (Card, 2023). In our study, we adopt a method loosely based on the latter approach, which we will elaborate on in Section 4.2.

2.3 Lexical Hate Speech Detection (LHSD)

Considering the potential application purposes of LSCD methods, addressing hate speech becomes a pressing concern, as neglecting changes in hateful meanings can lead to harmful consequences. While Hate Speech Detection (HSD) research has predominantly centered on identifying hate speech at the utterance level (Schmidt and Wiegand, 2017), a few works have addressed automatic detection at the lexical level, which is particularly relevant in the context of lexical change. Wiegand et al. (2018) presented an approach that utilizes a feature-based classification system to automatically expand a base lexicon of abusive words.

More recently, Hoeken et al. (2023) introduced a methodology for detecting lexical hate speech, involving the identification of a specific dimension within the embedding space of a language model that encodes hate. This dimension, estimated as the average difference vector of a set of lexical pairs that differ only with respect to the semantic dimension of hate, is then used to compare various word vectors. Using a pre-trained contextualized language model for generating lexical representation, this approach enables the prediction of hateful words within specific contexts.

2.4 Integrating HSD and LSCD

To our knowledge, the only contribution to the integration of hate speech detection and semantic change is done by McGillivray et al. (2022). Their study explores the feasibility of identifying offensive speech within data from 2020 using a model trained data from 2019. Their approach involves incorporating lexical semantic change scores as supplementary lexical features. Unlike our study, their primary focus is on contemporary hate speech detection and short-term meaning shifts. Nonetheless, their study illustrates the applicability of LSCD methods to a curated list of words that underwent shifts in offensive meanings. Still, the ability to filter out shifts that pertain solely to the semantic subdomain of hate remains unsolved.

A few other studies explore the shift of a specific dimension of meaning, that go beyond the predominant focus within LSCD on denotation. Charlesworth et al. (2022) employ static diachronic word embeddings trained on data reaching back to the 1800s (Hamilton et al., 2016) (in contrast to our contextualized LLM-based approach) and human-rated sentiment scores to explore to investigate how the social group representations and their perception have changed over time.

Another approach proposed by Basile et al. (2022) builds upon a ‘connotative hyperplane’ within embedding space, which is similar to the principle of an hate dimension. Shifts are quantified by measuring the difference in distances between word vectors and the hyperplane.

3 Data

3.1 Historical pamphlets as input data

Our study focuses on 16$^{\text{th}}$ century pamphlets, which provide a glimpse into conflicts and controversies associated with the Protestant Reformation in England and context-related language use. Pamphlets had been a new phenomenon in Early Modern England and were on the rise with the introduction of the printing press in the late 15$^{\text{th}}$ century. Much smaller, cheaper and faster in production than books at that time, pamphlets provided the opportunity to reach large audiences for the first time, which brought about a change in the dynamics of public debate (Dröse, 2021).

Although religious pamphlets came along in various shapes - poems, dialogues, sermons, treatises etc. - , a major shared characteristic is a polemical style in order to convince the readership of certain religious positions. Polemical language in the 16$^{\text{th}}$ century is described by historians and literary scholars as being persuasive, emotionally charged, and reactive (Almasy, 2008). The intention of Catholic and Protestant polemicists often was to argumentatively justify and demonstrate their sovereignty in interpreting religious issues. A major characteristic is a double audience (Steckel, 2018): not only were the pamphlets addressed at people sharing the same beliefs, but also at the respective opponents.

Thus, we find these texts riddled with derogatory language and hateful terms as we see in an illustrative statement made by Thomas Bell, an anti-Catholic author, in 1596, denoting Catholics as heretics: “the papistes are nothing else but flatte heretikes.” Moreover, the historical writings already reflect a sense of different nuances of hatefulness. For instance, in his *Actes and Monuments*, first published in 1563, the Protestant clergyman and writer John Foxe made a qualitative differentiation between hateful terms in a religio-political context: “I had rather be counted a king foolish and simple, then to be iudged a tiraunt or a seeker of bloude”. Hence, we can assume that hate speech constitutes a crucial feature of Early Modern English religious polemics and was subject to reflection at that time, too.

3.2 Period and text selection

Our data is sourced from Early English Books Online (EEBO)[2], an online database which provides the largest Early Modern English text corpus and includes publications from 1473 to 1700. Narrowing down the time frame to 1485-1603 allows us to look into possible changes in language use with the beginning of the English Reformation era. The texts were selected through an iterative keyword-based search, which ensured that they share the context of the Reformation. Appendix A lists the total set of keywords used. The data statistics of our final selection from EEBO are presented in Table 1.

The division into smaller periods of time is based

[2]https://www.proquest.com/eebo/index

Period	Phase	Texts	Sentences	Tokens
1485-1529	Catholic	14	31 692	852 823
1530-1552	Protestant	70	74 573	2 752 053
1553-1558	Catholic	20	24 846	809 885
1559-1579	Protestant	43	189 139	6 360 794
1580-1603	Protestant	162	477 896	16 768 865

Table 1: Statistics of texts per time period after final data selection.

on major political, societal and religious events and, as can be seen in Table 1, divided into periods of Catholic and Protestant monarchs: i. 1485-1529, ii. 1530-1552, iii. 1553-1558, iv. 1559-1579, and v. 1580-1603. The first phase marks the pre-reformation era under Henry VII. and Henry VIII (i.). With the 1530s, the Protestant Reformation in England gained momentum, Henry VIII. breaking with Rome and establishing Protestantism across England (ii.). After the reigns of Henry VIII. and Edward VI., Mary I. succeeded (iii.), who tried to re-establish the Catholic church. With Elizabeth I., a Protestant monarch followed again in 1558 (iv.). Anti-Catholic sentiments further increased and peaked during the 1580s (v.). Therefore, we can expect changes due to radicalization and changing political circumstances under which the texts were published.

For the present study, we chose to focus on only two of these time spans, taking into account 70 texts from 1530-1552 (ii.) and 162 texts from 1580-1603 (v.), in order to trace the diachronic change in Protestant polemical language. The difference in quantity aligns with the availability of publications, which continually increased from the beginning of the 16th century.

3.3 Cleaning and Normalization

Firstly, we removed both the header and footer sections, containing metadata, from the lowercased texts, along with the page numbers. Afterwards, we employed the sentence tokenizer provided by the Natural Language Toolkit (NLTK).

Initial analysis of the data showed significant spelling variations for identical words. Therefore we apply spelling normalization through a rule-based approach that generates a spelling dictionary which we apply to the whole corpus. A naive lookup technique like this showed most effective for historical text normalisation (in the case of in-vocabulary tokens) in the methodological evaluation conducted by Bollmann (2019). The details of our used method are specified in Appendix B.

4 Methods

4.1 Task and Procedure

In this paper we introduce an approach designed to tackle the task of **Lexical Semantic Hate Change Detection**, which we define as follows:

> Given a dataset D_0 from time period T_0, dataset D_1 from time period T_1, detect whether a target word gained or lost a hateful meaning between time T_0 and T_1.

Our approach ultimately yields a dataset with dual-aspect annotations: lexical semantic changes and lexical hatefulness. To capture potential changes of hateful meanings in our dataset (see Section 3), the selection of target words for the annotated dataset is guided by outcomes of both LHSD and LSCD methods. For both methods, we employ a historical BERT model, MacBERTh (Manjavacas Arevalo and Fonteyn, 2021)[3], which was trained on data spanning the years 1450 to 1950, also encompassing the EEBO database.

In the following, we present our method for LSCD (Section 4.2) and for LHSD (Section 4.3); a simple validation of LHSD on our historical data (Section 4.4). We also detail the manual annotation of lexical change and hatefulness (Section 4.5). The main idea of our approach is to first rank candidate words with respect to their semantic change and hatefulness score, and, based on the rankings, annotate a sample of potential target words to be able to evaluate the aumatic scoring.

4.2 LSCD

To measure changes in word meanings over time, we use a slightly simplified version of a recent methodology introduced by Card (2023). This method utilizes a BERT model's ability to predict masked words and involves the following steps for each target word. For a sample of contexts in which the target word occurs, we mask the target word and let the model predict its substitution. We gather the top 10 most probable substitutions for each instance (omitting stopwords, words with fewer than 3 characters or containing non-alphabetic characters). Across all target word instances, we calculate the frequency for each distinct substitute token, relative to the entire vocabulary of the model. Finally, the Jensen Shannon Divergence (JSD) is calculated

[3]We implemented the 'emanjavacas/MacBERTh' model using Hugging Face's *transformers* library (Wolf et al., 2020).

to quantify the difference in substitute frequency distributions between different time periods.

4.3 LHSD

To assess whether words carry a hateful connotation, we adopt the methodology introduced by Hoeken et al. (2023). Diverging from their approach, we apply it to a diachronic scenario. We create a hate dimension based on lexical pairs sourced from one time period. Subsequently, we project potential target terms from different time periods onto this dimension, allowing to determine the degree of hatefulness encoded in their representations and whether this has shifted over time.

Dimension creation. From the last time period (1580-1603), we create a set of lexical pairs of hateful terms and their neutral counterparts, i.e. terms referencing the same target group without any derogatory connotations. We extracted all unique nouns (using the Spacy library for POS tagging) from the texts in this period that occurred more than 10 times, ended with an ‘s’ (potentially targeting references to (groups of) people) and consist of more than 3 characters, resulting in a list of 5976 nouns. From this list, 65 potential hateful terms were selected for further analysis. An expert historian manually examined the contexts in which these terms were used and selected 10 terms that consistently demonstrated a highly hateful connotation across the majority of contexts in which they appeared. For these 10 terms, we identified their neutral counterparts, resulting in our set of lexical pairs as displayed in Table 2.

	Hateful term	**Neutral counterpart**
1	heretikes	protestants
2	hipocrites	catholikes
3	idolaters	catholikes
4	papists	catholikes
5	popelings	catholikes
6	traitours	catholikes
7	shavelings	monkes
8	harlots	women
9	strumpets	women
10	whores	women

Table 2: 10 pairs of hateful terms and their neutral counterparts, used for dimension creation, from the 1580-1603 dataset.

Following Hoeken et al. (2023), we computed a dimension vector as the mean distance vector of the set of lexical pairs. For every pair, an averaged lexical representation is generated across 10 contexts in which they occur. We manually selected the contexts for each term ensuring that each context distinctly represents a hateful word as hateful and a neutral counterpart as neutral. This also guarantees that both parts of the lexical pair refer to the same entity, fulfilling the requirement of a difference, solely concerning the hateful dimension, between the two. We employed the MacBERTh model to extract each contextualized representation by averaging over all the hidden layers and the sentence positions of the subwords forming the pair.

Dimension projection. For a contextualized representation of a target word, the degree of hate encoded in it can be determined by projecting it on the hate dimension. This is established by computing the cosine distance between the two vectors. Positive angle values indicate a hateful connotation, while negative values do not.

4.4 Identifying historical hateful terms

To assess the applicability of the above-mentioned method for detecting lexical hate speech, originally devised for synchronic use, in the context of historical and diachronic data, we conduct a proof-of-concept validation analysis.

For the two periods under investigation, we extracted a list of terms adhering to the same criteria as those further employed throughout this study[4]. This yielded 1490 terms from the period 1530-1552 and 6338 terms from 1580-1603. Subsequently, we applied the hate projection method to 100 contextual representations of each noun, or fewer if a word occurred fewer than 100 times.

In Table 3, we present the top 25 words from each period, ranked by their average projection values, indicative of the degree of hate encoded in their representations. A historian further evaluated the hatefulness of these words, drawing on their historical expertise, historical dictionaries, or examination of the contexts in which the words occurred. The majority of these words (all but one to three per period) were confirmed to convey hateful meanings. This implies that, given a small sample of known hateful terms to create a dimension vector, this method can effectively detect hateful terms in *different* historical periods based on a small sample of known terms from *one* period.

[4] i.e. nouns occurring more than 10 times, ending with ‘s’ and consisting of more than 3 characters.

1530-1552	1580-1603
extorcioners, liars, liers, buggerers, idolatres, stubburnes, fals, abusions, aulters, dregges, blasphemers, baudes, bablinges, *gobbettes*, deuelles, mischefes, idolatours, deuels, robbers, wrincles, sclaunders, persecutours, sorcerers, idolles, vnthankefulnes	liars, abhominations, inchaunters, *diotrephes*, libidinis, hipocrits, iuglers, backbiters, corrupters, impostures, extortioners, liers, iarres, whoredomes, puddles, lascivious, vilanies, bawdes, *iambres*, fornications, varlets, abusers, baudes, *paunches*, iuglings

Table 3: Top 25 words with highest average projection values in 1530-1552 and 1580-1603. The hatefulness of all but *italic* words were confirmed by an historian expert.

4.5 Annotation

4.5.1 Target word selection

To scale the validation of our approach sketched in Section 4.4, we select a larger set of words for annotation of lexical change and hatefulness. From the intersection of the vocabularies of D_0 and D_1 (corresponding to the data from 1530-1552 (T_0) and 1580-1603 (T_1) respectively), all nouns that fits to the selection criteria and not used for dimension creation are extracted, resulting in 1163 nouns.

For each of these nouns, we randomly extract up to 100 contexts per period. Then, both the LHSD method as well as the LSCD method are applied on all instances, as explained in Section 4.4. As a result, we obtain for each noun, one semantic change value and two projection values (reflecting their predicted hatefulness in each period). The difference between the two projection values is computed for each word to calculate the "hate change" score.

For the creation of the pilot dataset, a selection of 100 nouns (target words) is made. This selection includes the top 20 and bottom 20 words ranked by their semantic change value as well as the top 20 and bottom 20 words ranked by their hate change score. The resulting sets can be found in Appendix C. Additionally, we randomly 20 sample nouns to end up with a total set size of 100.

4.5.2 Annotation scheme & procedure

The annotation study serves two primary objectives: 1) publishing a dataset with rich annotations, and 2) providing a test-set for the computational approaches employed. For annotation we predominantly adopt the Diachronic Usage Relatedness (DURel) framework by Schlechtweg et al. (2018) that is designed for annotating lexical semantic changes. We extend this framework by incorporating annotations of hatefulness.

For each of the 100 target words, 10 contexts are randomly selected from each time period. From this set of 20 contexts, we randomly select 10 pairs of contexts either from the same period or from different ones. Consequently, the final test-set comprises a total of 1000 pair instances. For each text pair with a highlighted word, annotators are asked to evaluate the lexical semantic change and hatefulness. An example of an annotation instance is provided in Appendix C.

To annotate lexical semantic change, we employ the 4-point scale of relatedness as presented in the DURel framework. For the annotation of hatefulness we adopt the three-class scheme of Vigna et al. (2017), and add 'Cannot decide' to it, see Table 4.

4	Identical		
3	Closely related	**2**	Strongly hateful
2	Distantly related	**1**	Weakly hateful
1	Unrelated	**0**	Not hateful
-	Cannot decide	**-**	Cannot decide

Table 4: Four-level scale of semantic relatedness (Schlechtweg et al., 2018) (left) and three-level scale of hatefulness (Vigna et al., 2017) (right)

The annotations have been performed by two experts on medieval and early modern history. Both annotators were provided with the same instructions and illustrative examples[5]. For reasons of feasibility, the second annotator undertook the annotation of a subset of the data, encompassing half (50) of the target words, each with the same set of 10 sentence pairs as in the complete dataset. The subset retained the same distribution with respect to high and low JSD and projection difference values.

5 Results

5.1 Annotation outcomes

Agreement. We analyze the agreement of our two annotators[6] and report the inter-annotator agreement in Table 5. Both for semantic relatedness, which involves all sentence pairs rated by both annotators (total of 435), as well as the annotation of hatefulness, which involves all individual sentences rated by both annotators (total of 870), show a fair agreement in terms of Cohen's Kappa (0.247 and 0.315, respectively).

Semantic change. To transform the human annotations of semantic relatedness between pairs of sentences (from the same or different time periods) into values that indicate the semantic change

[5]The annotation instructions can be accessed on our GitHub repository.

[6]'Cannot decide' annotations are omitted. The first annotator flagged 88 out of 1000 instances with one or more 'Cannot decide'. For the second annotator this was 15 out of 500.

	Sem. rel (n = 435)	Hate (n = 870)
Cohen's κ	0.247	0.315
Pearson's r	0.576*	0.511*

Table 5: Inter-annotator agreement for the two annotators on their ratings of semantic relatedness and hatefulness (* = significant).

of target words between the two time periods we compute the COMPARE score. This score, also introduced within the DURel framework, is defined as the average between sentence pairs from different periods (Schlechtweg et al., 2018). To facilitate a more intuitive and straightforward analysis, we convert the scaled human ratings into binary values by applying boundary thresholds. For the COMPARE scores, any score below 4 is interpreted as change whereas a score of 4 as no change.

Hatefulness. In contrast to the change scores, which are analyzed on type level only, i.e. one aggregated result value for each target word, the hatefulness scores are also analyzed on token level, involving all unique sentence ratings from both time periods. For transformation to a binary classification of each target word, any *average* hate rating greater than 0 is interpreted as hateful, and 0 as not hateful.

Changes of hateful meanings. Combining the binary outcomes for semantic change and hatefulness annotations, allows to distinguish words that are (on average) classified as both hateful and having undergone semantic changes from those that are not. Table 6 reports the number of words on categorized as changed in meaning, conveying a hateful meaning and those falling into both categories simultaneously. Overall, we obtain 23 types that changed their meaning wrt. hatefulness, yet the distribution also indicates the challenging nature of the task that can be attributed to the sparseness of this case.

	Annotator 1 (all)	1 (n=50)	2 (n=50)
changed	26	14	31
hateful	13	7	35
hateful + changed	8	3	23
Out of	99	50	50

Table 6: Number of target words whose meaning is on average classified as changed, hateful, and both by the different annotators, n = number of observation

5.2 Methods evaluation

We leverage the created (pilot) dataset enriched with two-aspect annotations to evaluate the outcomes of the proposed computational methods.

Semantic change. When comparing the JSD values from the LSCD method with the human change scores (as previously explained), we expect a negative correlation, as higher JSD values indicate higher difference between time periods while lower human scores indicate the same. To transform the continuous JSD values into binary classes, we set the mean JSD value across all words considered for comparison as the threshold between change and no change (following the common practice in Schlechtweg et al. (2020)).

Hatefulness. The hate dimension method produced projection values between -1 and 1, for each contextualized instance of a target word. We compare these output values with the human hatefulness ratings for all unique sentences. For binary classification, all positive values are interpreted as hateful, and negative ones as not hateful.

Annotator	Semantic change binary	graded	n	Hatefulness binary	graded	n
1 (all)	0.61	-0.39	99	0.66	0.21	1297
1	0.68	-0.43	50	0.61	0.26	683
2	0.74	-0.54	50	0.62	0.28	687
Avg.	0.74	-0.52	50	0.62	0.33	683

Table 7: Pearson's r for graded and accuracy for binary outcomes of computational approaches compared with human annotations; n = number of observations; all evaluation values are significant.

Table 7 presents the results of the graded evaluation using the Pearson correlation, while for the binary classification, we provide the accuracy scores. To determine significance for the latter, we employ the chi-squared test. We report the evaluation scores for each annotator individually (1 and 2), as well as aggregated by comparing them with the average ratings provided by both annotators[7].

Overall, both the accuracy scores (ranging between 0.61 and 0.74) and the correlation scores (between 0.26 and 0.52), indicate moderate performance of the computational approaches. These results align with the the inherent complexity of the tasks as demonstrated by the fair inter-annotator

[7]We computed Pearson correlation and significance tests using the SciPy library and Cohen's Kappa and the classification report using the scikit-learn library, for Python

agreement. Furthermore, the task of predicting hatefulness yields lower scores compared to predicting semantic change, which implies a difference in the complexity of the two tasks, with the former being more complex than the latter.

Changes of hateful meanings. Similarly to the human annotations, we merge the binary outcomes from the two computational approaches. This enables us to evaluate the classification of words being both hateful and undergoing changes in meaning. In Table 8 the classification by computational approaches is compared with the human annotation outcomes, averaged across the two annotators.

	prec.	recall	F1	n
hateful + changed	0.73	0.42	0.53	19
not hateful + changed	0.72	0.90	0.80	31
macro avg	0.72	0.66	0.67	50

Table 8: Report on the classification of change of hateful meanings compared with average annotator outcomes.

Unsurprisingly, the performance on the combined tasks demonstrates a trade-off between the individual task performances as reported in Table 7. The results reveal that our methods accurately identify around half of the words categorized as shifting in hateful meanings. The low recall rate indicates that false negatives constitute the predominant error type.

5.3 Error analysis

Overall, a potential explanation for the discrepancy between the human annotations and LSCD method predictions (not the LHSD method) might be attributed to the fact that human annotators were tasked with rating an average of approximately 10 contexts per time frame for each target word, whereas the method outcomes derived their predictions from a sample of up to 100 contexts.

To gain a deeper understanding of the specific errors made by the methods we conducted a manual analysis of error cases demonstrated in the comparisons between the methods and *both* annotators. These cases concerned all error types, except for non-existing false negatives of semantic change detection.

Semantic change: false positives. Discrepancies in the detection of semantic change between the computational method and human annotations do not necessarily imply a failure of the method, but could be due to annotation granularity, with the target word ‘swearers’ being an example case. The method’s subtle change detection might not align with the expert annotations differentiating only between “weakly" and “strongly hateful". Consequently, the erroneous detection of *semantic change* leads to ‘duns’ and ‘swearers’ being false positives for the classification of *change in hateful meanings*, too.

Hatefulness: false negatives. A potential reason for this error type is usage of metaphor. For instance, ‘foxes’ was frequently used by Protestants to refer to their opponents in a hateful manner, examplified by a statement made by Andrew Willet in 1592: “They are the foxes that destroy the lords vineyard.” (For a deeper analysis of metaphors used in polemical Reformation writings, see Kelly (2015)). This consequently led to ‘foxes’ being a false negative in the classification of *change in hateful meaning*, too.

Hatefulness: false positives. The target words falsely detected as hateful by the method are: ‘anselmus’, ‘higinus’, ‘nauclerus’, ‘sigebertus’. These all are names which do not carry hateful meanings themselves but predominantly occur within hateful contexts, which potentially leads our method to predict a hateful meaning.

6 Discussion & Conclusion

Our study introduces the novel task of detecting changes of hateful word meanings in historical texts. Our interdisciplinary approach combines Lexical Semantic Change Detection and Hate Speech Detection. We leverage historical expertise to generate a pilot dataset with two-aspect annotations, a valuable resource for the evaluation of computational methods. While our methods showed effective precision in detecting hateful words that changed their meaning throughout the 16$^{\text{th}}$ century, they also underscored the complexity of (the combination of) the tasks, as evidenced by the human interrater agreement scores.

The exploration of hate speech within historical discourse poses particular challenges. Most importantly, we acknowledge inherent limitations as we may never achieve a perfect reflection of historical connotations. Still, our framework aims at a closer grasp of the past by combining historical research and linguistic analysis. The bounded text selection and the limited annotated data (for reasons of feasibility) pose challenges to the robustness and generalizability of our findings, pertaining to the efficacy of the employed methods as well as the outcomes we have presented. Therefore, our conclusions should be further validated in follow-up research that incorporates more diverse textual sources and enhances the quantity (and quality) of the annotated data. We further propose to broaden the scope beyond nouns, as verbs and adverbial phrases can also convey hate in the form of devaluation of action. Moreover, our error analyses highlighted the prevalence of metaphors for expressing hateful meanings, suggesting another direction further research. Finally, expanding the focus to longer time-spans or conducting cross-language comparisons could also yield valuable insights.

In conclusion, our paper lays foundations for advancing the analysis of lexical change of a specific domain in historical data. Particularly, our interdisciplinary framework paves the way to an expanded dataset and the development of better computational methods for detecting the evolution of historical hate speech.

Limitations

Going beyond the reflection of our work in 6, we would like to further point to some methodological limitations in our study. Firstly, the decision for the used sentence split method appeared not well-suited for digitized historical texts, with punctuation to indicate sentence breaks often missing. This resulted in some flawed sentences, thereby providing limited context information as input for the model's predictions. For further research we would therefore either opt for manual sentence splitting or better trained sentence split algorithms for historical data.

Additionally, the employed spelling normalization method fails to encompass all possible variations, potentially resulting in overlooked or misinterpreted semantic changes that could be perceived as errors. For instance, the word 'sees', which in both time periods could denote 'seas', referring to the ocean; whereas in the later period, it was also utilized in the context of 'bishop's sees', referring to their realm of power. In this case, a gain in word meaning is wrongly identified as the secondary meaning already existed in the earlier period, albeit in the orthographic variant 'sedes'.

Lastly, the method also catches target words if they are part of another word: e.g. the target word 'gaines' also occurs as part of the word 'gainesayers'. Therefore, sentences mentioning both words are taken into account, while we are only interested in the former.

Ethics Statement

Investigating hate speech brings about ethical issues to reflect upon. Unlike modern data typically used for HSD, the textual data from the 16th century we are drawing on is publicly available along with metadata, such as the authors' names. There is no need for anonymization. On the contrary, it is of high value to be able to access context information to further work with the results of our method once it is fully developed. However, we are aware that filtering out hate speech, in our case hateful terms particularly used against Catholics, allows for reproduction in modern days, especially because we still face Anti-Catholicism in present-day societies. Therefore, it is crucial, also for future work, to ensure that the methods' results are always viewed with regards to their historical context and only used for improving NLP methods in order to detect and potentially avoid further usage of hate speech or as a data basis for historical and cultural studies.

Acknowledgements

The authors acknowledge financial support by the project "SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), which is funded by the program "Netzwerke 2021" of the Ministry of Culture and Science of the State of Northrhine Westphalia, Germany.

Additionally, we would like to thank Melvin Wilde for his annotations, expertise and support.

References

Rudolph P. Almasy. 2008. *Rhetoric and Apologetics*, pages 121 – 150. Brill, Leiden.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. Diacrita@ evalita2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task. *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

Valerio Basile, Tommaso Caselli, Anna Koufakou, and Viviana Patti. 2022. Automatically computing connotative shifts of lexical items. In *Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia,* *Spain, June 15–17, 2022, Proceedings*, pages 425–436. Springer.

Elisa Bassignana, Valerio Basile, Viviana Patti, et al. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings*, volume 2253, pages 1–6. CEUR-WS.

Antoinina Bevan Zlatar. 2011. *Reformation Fictions. Polemical Protestant Dialogues in Elizabethan England*. Oxford University Press.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. 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 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.

Dallas Card. 2023. Substitution-based semantic change detection using contextual embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 590–602, Toronto, Canada. Association for Computational Linguistics.

Tessa E. S. Charlesworth, Aylin Caliskan, and Mahzarin R. Banaji. 2022. Historical representations of social groups across 200 years of word embeddings from google books. *Proceedings of the National Academy of Sciences*, 119(28):e2121798119.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Albrecht Dröse. 2021. Invektive Affordanzen der Kommunikationsform Flugschrift. *Kulturwissenschaftliche Zeitschrift*, 6(1):37–62.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Sanne Hoeken, Sina Zarrieß, and Ozge Alacam. 2023. Identifying slurs and lexical hate speech via lightweight dimension projection in embedding space. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 278–289, Toronto, Canada. Association for Computational Linguistics.

Erin Katherine Kelly. 2015. Chasing the fox and the wolf. Hunting in the religious polemic of William Turner. *Reformation*, 20(2):113–125.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

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

Barbara McGillivray, Malithi Alahapperuma, Jonathan Cook, Chiara Di Bonaventura, Albert Meroño-Peñuela, Gareth Tyson, and Steven Wilson. 2022. Leveraging time-dependent lexical features for offensive language detection. In *Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP)*, pages 39–54, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection. *arXiv preprint arXiv:2304.01666*.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Maxim Rachinskiy and Nikolay Arefyev. 2022. GlossReader at LSCDiscovery: Train to select a proper gloss in English – discover lexical semantic change in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 198–203, Dublin, Ireland. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Silke Schwandt. 2018. Digitale Methoden für die Historische Semantik. Auf den Spuren von Begriffen in digitalen Korpora. *Geschichte und Gesellschaft*, 44(1):107–134.

Gerd Schwerhoff. 2020. Invektivität und geschichtswissenschaft konstellationen der herabsetzung in historischer perspektive. ein forschungskonzept. *Historische Zeitschrift*, 311(1):1–36.

Sita Steckel. 2018. Verging on the polemical. Towards an interdisciplinary approach to medieval religious polemic. *Medieval Worlds*, 7(1):2–60.

Nina Tahmasebia, Lars Borina, and Adam Jatowtb. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6:1.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Italian Conference on Cybersecurity*.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

A Keywords used for text selection

The corpus of texts we used as input data was generated through an iterative keyword-based search using the following keywords.

First selection: catholic*, church, faith*, invective*, libel*, protestant*, pamphlet*, pope, religio*, reformatio*, reformer*, religio*. (Asterisks represent wildcards in the search mask.)

Second selection: harlots, heretics, hypocrites, papists, strumpets, whores.

B Spelling normalization

We generated a substitution dictionary, tailored to our dataset, aiming to transform word forms into their most modern spelling variations within the corpus (e.g. transforming 'shauelyngs' and 'shauelings' to 'shavelings'). To achieve this, we created a set of rules for character substitutions, grounded in regular expressions. We applied these rules to all words in the vocabulary of the raw data collection. If a substitution resulted in an existing word in the (same) vocabulary, we included the before-and-after substitution pair in the dictionary. An overview of these rules and the corresponding procedure (in code) are presented below. We applied the mappings to the entire corpus.

```
import re

def get_variant(word):
  word = word.replace('ā', 'an')
  word = word.replace('ū', 'un')
  word = word.replace('ē', 'en')
  word = word.replace('ā', 'am')
  word = word.replace('ū', 'um')
  word = word.replace('ē', 'em')
  word = re.sub("uy", r"vi", word)
  word = re.sub("([^q])u([aeiou])", r"\1
      v\2", word)
  word = word.replace('vv', 'w')
  word = re.sub(r"^vh", "wh", word)
  word = re.sub(r"v([bgnprstx])", r"u\1"
      , word)
  word = re.sub(r"y", "i", word) if word
      != "i" else word
  word = re.sub(r"ie$", "y", word)
  word = re.sub("([aeiou])ie", r"\1y",
      word)
  word = re.sub(r"i$", "y", word) if
      word != "i" else word
  word = re.sub(r"^iou", "you", word)
  return word

for w in vocab:
  if w != get_variant(w) and get_variant
      (w) in vocab:
    dictionary[w] = get_modern_variant(w
        )
```

C Target words for test set

Method outcomes for hate & semantic change to guide target word selection. (Random sample is not included here)

Hate changes.

- Top 20 projection value differences (neutral to hateful) between 1530-1553 and 1580-1603: counsailours, abbayes, tailes, higinus, dainties, swearers, hornes, adonias, winchesters, founders, notes, autours, sins, ananias, pastoures, agnus, adversaries, ensamples, heremites, duns

- Bottom 20 projection value differences (hateful to neutral) between 1530-1553 and 1580-1603: dedes, honours, affections, companies, purenes, freres, theues, affectes, cerimonies, businesses, evilles, noes, sclaunders, fabianus, luthers, holines, fees, plays, lordshippes, fines

Semantic changes.

- Top 20 JSD values (most changed between 1530-1553 and 1580-1603): strokes, males, dainties, winchesters, provisions, doctores, gaines, hominibus, affectes, womens, accountes, foxes, bargaines, parsons, giles, strengthes, wais, faculties, sees, professions

- Bottom 20 JSD values (most stable between 1530-1553 and 1580-1603): dionisius, preestes, presbiteros, aulters, berengarius, galathians, otherwhiles, polidorus, anselmus, rechabites, lanfrancus, ciprianus, sigebertus, apocalips, cauillations, ezechias, nauclerus, fulgentius, chrisostoms

Semantic & hate changes.

- Intersection of Top 20 projection value differences (neutral to hateful) and Top 20 JSD values (most changed): dainties, winchesters

- Intersection of Bottom 20 projection value differences (hateful to neutral) and Top 20 JSD values (most changed): affectes

Figure 2 displays an example of an annotation instance.

D Annotation example

target	sentence1	sentence2	semantic relatedness	hate1	hate2
swearers	for if there be a god, as i am certenly persuaded ther is, i am sure that these abhominable swearers shall not escape vnponysshed, let then esteme their sinne as light & as litle as they list, yea i am sure, i e vengeaunce of god hangeth over their heades, wher so ever they be.	the lorde will not holde him gyltelesse that taketh his name in vayne. let not these swearers therfore glory in their wickednes, and thinke i • they shall escape vnponished, because god takethe not vengeaunce on them streight ways , but rather let them thincke that their damnaciō shall be so muche the more greuous, seing they escape so longe without punishment.	4	2	2

Figure 2: Example of annotation instance