

# Chronoberg: A 250-Year Diachronic Corpus with Temporal Affective Lexicons for Language Model Analysis

Niharika Hegde<sup>1,4\*</sup>, Subarnaduti Paul<sup>2\*</sup>, Lars Joel-Frey<sup>4</sup>, Manuel Brack<sup>1,3‡</sup>,  
Kristian Kersting<sup>1,3,4,5</sup>, Martin Mundt<sup>2†</sup>, Patrick Schramowski<sup>1,3,4,6†</sup>

<sup>1</sup>DFKI, <sup>2</sup>University of Bremen, <sup>3</sup>hessian.AI,

<sup>4</sup>TU Darmstadt, <sup>5</sup>Centre for Cognitive Science, TU Darmstadt, <sup>6</sup>CERTAIN

Correspondence: [niharika.hegde@dfki.de](mailto:niharika.hegde@dfki.de), [spaul@uni-bremen.de](mailto:spaul@uni-bremen.de)

## Abstract

Large language models (LLMs) excel at operating at scale by leveraging social media and data crawled from the web. Whereas existing corpora are diverse, their frequent lack of long-term temporal structure may however limit an LLM’s ability to contextualize semantic and normative evolution of language and to capture diachronic variation. To support the latter, we introduce Chronoberg<sup>1</sup>, a temporally structured corpus of English book texts spanning 250 years, curated from Project Gutenberg and enriched with a variety of temporal annotations. First, the edited nature of books enables us to quantify lexical semantic change through time-sensitive Valence-Arousal-Dominance analysis and to support historically-grounded interpretation. With the lexicons at hand, we demonstrate a need for modern LLM-based tools to better situate their contextualization of sentiment and detection of discriminatory language across various time-periods. In fact, we show how these tools struggle with diachronic shifts in meaning, emphasizing the need for temporally aware training and evaluation pipelines, and positioning Chronoberg as a scalable resource for respective studies.

**Disclaimer:** This paper includes language and display of samples that could be offensive to readers.

## 1 Introduction

Language evolves continuously, reflecting shifts in knowledge, culture, and social norms. Yet, most large language models (LLMs) are trained on near-stationary datasets. Existing web-crawled corpora such as Common Crawl and Wikipedia best feature short-horizon temporal variations. Alas, without

temporal grounding, language models risk conflating historical and contemporary meanings, for example, misinterpreting phrases such as “*Where is the woman to strew the flowers?*” by applying modern connotations. Such misreadings can distort semantic understanding, but can also amplify outdated stereotypes and ethical blind spots (Blodgett et al., 2020). As language and societal norms continue to evolve, it becomes increasingly critical to understand how models adapt their representations to ongoing and possible future linguistic change (Dhingra et al., 2022).

This challenge is complex, as these shifts operate over centuries, not months. Semantic change ranges from evaluative shifts in affective polarity (“pejoration”, “amelioration”) to non-evaluative shifts in meaning scope (“metaphor”, “narrowing”, “generalization”) (Jeffers and Lehist, 1979; Traugott, 2017). Existing resources (Google Books; Davies, 2015, 2012; Jang et al., 2022) do not simultaneously provide the scale, temporal depth, and structured annotation needed to study them, or analyse where modern classifiers fail because of them. For such temporal contextualization, collections of edited books as a form of curated archive provide a more suitable resource, as demonstrated by pioneering “culturomics” studies (Michel et al., 2011). However, no existing book-derived resources simultaneously satisfy the above requirements, while also being openly accessible. Large n-gram archives (Michel et al., 2011; ProQuest, 2008) discard sentence context; curated corpora (Davies, 2012; Alatrash et al., 2020) are too small or remain closed-source; and recent temporal datasets (Li et al., 2025; Jang et al., 2022) primarily target short-horizon factual drift.

To bridge this gap, we introduce Chronoberg, a diachronic corpus of 2.7 billion tokens spanning 250 years of English literary texts (1750–2000), curated from Project Gutenberg (Project Gutenberg) and enriched with temporal annota-

<sup>\*</sup>Equal contribution. <sup>†</sup>Equal supervision. <sup>‡</sup>Work done while at DFKI and TU Darmstadt.

<sup>1</sup>Chronoberg is publicly available on HuggingFace at <https://huggingface.co/datasets/spaul25/Chronoberg>, with code at <https://github.com/paulsubarna/Chronoberg>.

tions. Three contributions are central. First, we construct temporally calibrated Valence-Arousal-Dominance (VAD) lexicons covering over 300,000 words across five 50-year epochs by propagating human-annotated VAD scores (Mohammad, 2018). Together with sentence-level VAD annotations for the full corpus, these lexicons enable fine-grained tracking of sentiment trends across time. Second, we validate the lexicons against words with documented evaluative and non-evaluative changes from the linguistics literature (Jeffers and Lehiste, 1979; Hamilton et al., 2016; Schlechtweg et al., 2020), showing that evaluative shifts with clear VAD signal are correctly detected. Third, we demonstrate a practical consequence of temporal misalignment: contemporary LLM-based hate-speech classifiers (Vidgen et al., 2021; Lees et al., 2022) conflate modern connotations with historical reality and how contemporary language models perform under temporal shift. Chronoberg, including all annotations, will be made publicly available, supporting future work in diachronic lexical analysis and temporally aware machine learning and evaluation.

## 2 Chronoberg Dataset

We describe the construction of Chronoberg, a temporal corpus designed to study the semantic evolution of the English language across centuries. We first situate it within existing resources, before detailing the pipeline, depicted in Figure 1, covering curation, metadata inference, preprocessing, filtering, and VAD lexicon creation.

### 2.1 Relation to Existing Works

Existing temporal corpora fall short in different ways. Large n-gram archives such as Google Books and Early English Books (EEBO) (Michel et al., 2011; ProQuest, 2008; Text Creation Partnership, 2008) offer broad historical coverage but discard sentence-level context, whereas curated corpora such as COHA (Davies, 2015), COCA (Davies, 2012) and CCOHA (Alatrash et al., 2020) provide richer annotations but are not freely accessible at scale. Similarly, more recent works in temporal dataset construction (Li et al., 2025; Jang et al., 2022) only target short-horizon factual drift, leaving long-horizon connotative evolution largely understudied computationally, despite a rich, extensive theoretical grounding in historical linguistics.

On the theoretical side, historical linguistics has established a principled typology of mean-

ing change (Bloomfield, 1933; Jeffers and Lehiste, 1979; Lehrer, 1985; Traugott and Dasher, 2001), documented in resources such as the Historical Thesaurus of the OED (Kay et al., 2009). Computationally, distributional methods detect semantic shift at scale (Gulordava and Baroni, 2011; Jatowt and Duh, 2014; Kulkarni et al., 2015), culminating in word embedding approaches that reveal statistical laws of change (Hamilton et al., 2016) and transformer models that resolve individual senses over time (Schlechtweg et al., 2020; Edmondson, 2022; Cassotti et al., 2024).

These methods characterise whether and how meanings shift, but not the affective direction of that shift. This affective dimension is captured by Valence-Arousal-Dominance (VAD) frameworks, which encode polarity, activation level, and perceived control (Osgood et al., 1957; Russell, 2003), extended to tens of thousands of words through human annotation (Mohammad, 2018, 2025). These lexicons, however, are synchronic by construction, reflecting contemporary usage only.

This is consequential because, without connecting semantic shift to affective direction, downstream tasks such as hate-speech detection, sentiment analysis, and social bias auditing cannot account for the fact that the same word may carry different charge in a different era. This limitation is compounded by evidence that shifting social attitudes become embedded in NLP systems (Mendelsohn et al., 2020; Queerinaï et al., 2023) and that rapidly evolving language undermines classifier robustness (Keidar et al., 2022; Mehta and Giunchiglia, 2025). Chronoberg addresses these gaps by combining full-text coverage, yearly temporal granularity, long-horizon historical depth, and structured affective annotation over a 250-year period.

### 2.2 Collection Methodology

We build Chronoberg on (Project Gutenberg), an openly accessible corpus of literary texts that provides copyright-free English works at scale, in plain text and HTML formats, with metadata accessible via an API. However, the metadata is not directly suitable for temporal analysis. Original publication dates are frequently absent, incomplete, or inaccurate; available dates often reflect digitization, cataloguing, or release events rather than first publication. This poses a significant challenge for diachronic analysis, since assigning texts to incorrect years would directly

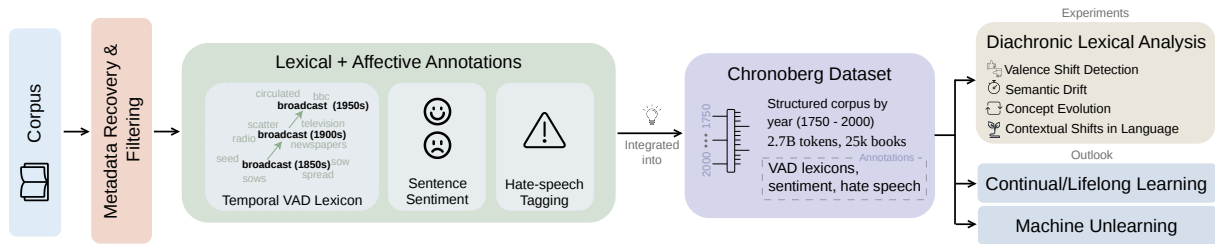


Figure 1: Overview of the Chronoberg dataset pipeline, spanning corpus curation, metadata recovery, and diachronic lexical analysis. The resulting annotations, including VAD lexicons and sentiment scores, form an integral part of the dataset and support downstream machine learning investigation.

distort estimates of semantic and affective change.

**Metadata Recovery.** We developed an inference pipeline that combines internal metadata with queries to external bibliographic sources (OpenLibrary and Wikipedia) and checks the consistency of inferred years against authors’ lifespans. Manual verification of 100 randomly sampled books (spanning years 1611-1912) identified OpenLibrary as the most reliable source, achieving a mean absolute error (MAE) of  $\pm 3.05$  years and standard deviation (SD) of 5.20 years (see Appendix A.3). Although this yields an uncertainty of 3-5 years, the margin is acceptable since our analysis operates at the scale of decades rather than years, comfortably exceeding the variance. We accordingly recommend that future studies adopt temporal bins no smaller than 15 years.

**Filtering.** Using OpenLibrary as the default inference source based on the prior validation results, we excluded books without an inferrable publication year or known author, discarded posthumous editions via author lifespan checks, and retained only English-language texts published between 1750 and 2000. The resulting chronological backbone of Chronoberg contains 25,061 out of 73,500 books in Project Gutenberg, allowing year-by-year aggregation, annotation, and subsequent analysis.

### 2.3 Lexical and Affective Annotations

We next describe the construction of Chronoberg’s core analytical contribution: a set of temporally aligned lexicons that tracks semantic change in word meanings across the 250 years. Methodologies for studying semantic change have evolved from qualitative, manual linguistic analysis (Bréal, 1897; Ullmann, 1962) toward large-scale distributional approaches, with alignment techniques such as PPMI (Bullinaria and Levy, 2007), Word2Vec

(Mikolov et al., 2013), SVD (Levy et al., 2015) and CADE (Di Carlo et al., 2019) enabling reliable recovery of directional semantic shifts that experts can verify (Blank, 1997; Kulkarni et al., 2015; Hamilton et al., 2016; Schlechtweg et al., 2020; Cassotti et al., 2024). These methods, however, only characterise whether and how meanings shift, but not the affective direction of that shift.

Influential factor analysis (Osgood et al., 1957; Russell, 2003) addresses this by decomposing affective meaning into three interpretable dimensions: Valence, Arousal, and Dominance, with human-annotated VAD lexicons extending this framework to 45,000 contemporary English words (Mohammad, 2018, 2025). These lexicons, however, are synchronic by construction and cannot track the shifting connotations of words (e.g., *broadcast* or *febrile*) whose historical usage modern ratings systematically misrepresent (Perc, 2012). In Chronoberg, we address this by combining diachronic semantics with human-annotated VAD scoring, proposing a novel set of temporally aligned lexicons spanning over 300,000 words across five 50-year epochs alongside sentence-level VAD scores for the full corpus.

**Temporal VAD Lexicons.** We model the shifts in word meaning using diachronic distributional semantics, learning a high-dimensional vector for each word from its co-occurrence patterns within a given time period.

Concretely, we train separate Word2Vec (Mikolov et al., 2013) models on a temporal slice of the Chronoberg corpus, with each slice corresponding to a 50-year interval between 1750 and 2000, chosen according to the observed variance in publication data estimation and including a reasonable safety buffer. Following validation to select suitable hyperparameters, training for each interval has been conducted for 10 epochs with a context

window size of 5 tokens on either side and an embedding vector dimensionality of 300. We note that interval boundaries do not correspond to any specific historical eras or expert-defined periodization, and leave these to future work. The resulting embedding spaces are subsequently aligned using Compass Aligned Distributional Embeddings (CADE) (Di Carlo et al., 2019), which facilitates the direct comparison of word vectors across different decades by anchoring them to a shared compass model of the full corpus.

We chose Word2Vec + CADE for being computationally efficient over 250 years of data and reliable for diachronic nearest-neighbour retrieval. This aligns with our VAD propagation method, which requires stable vector-space alignment to preserve local geometric consistency for cross-temporal semantic neighbourhood retrieval, a property that CADE’s compass-based alignment is specifically designed to ensure. While transformer embeddings (Devlin et al., 2019) could offer richer contextual representations, we do not expect fundamental shifts in VAD interpretation, as the core compression principles remain unchanged.

We then estimate VAD scores for each target ( $w$ ) by selecting the Top-K nearest neighbours ( $\mathcal{N}_k(w)$ ) in the embedding space and averaging their corresponding VAD values from the human-annotated NRC VAD lexicon (Mohammad, 2018, 2025):

$$\begin{aligned} \mathcal{N}_k(w) &= \text{Top}k_{u \in \mathcal{V}_{\text{VAD}} \setminus \{w\}} s(e_w, e_u), \\ \hat{\mathbf{A}}_{\text{VAD}}(w) &= \frac{1}{|\mathcal{N}_k(w)|} \sum_{u \in \mathcal{N}_k(w)} \mathbf{A}_{\text{VAD}}(u). \end{aligned} \quad (1)$$

Here,  $e_w \in \mathbb{R}^d$  is the embedding of target word  $w$ ,  $s(e_w, e_u)$  is the cosine similarity, and  $\mathcal{V}_{\text{VAD}}$  the set of words with NRC VAD annotations  $\mathbf{A}_{\text{VAD}}(u) \in \mathbb{R}^3$ ; the target word itself is excluded from its own neighbourhood set during propagation. We thus make use of the human-annotated scores in the NRC VAD lexicon, which primarily reflect contemporary interpretations of words, but re-contextualize them computationally to account for historical contexts or diachronic semantic shifts that certain words may have undergone. To this end, we use modern valence scores as anchors to detect relative semantic drift. While this may introduce systematic anachronisms, it is unavoidable and presents a natural choice, as retrospective human annotations at scale from several hundreds of years ago are practically unattainable, and few, if any, historians can operate at such a scale. For our

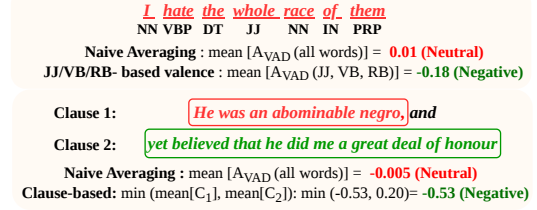


Figure 2: Sentence-level VAD scoring in Chronoberg. Instead of simple averaging across all words, we introduce a two-stage aggregation strategy: (i) part-of-speech-based averaging over adjectives, verbs, and adverbs (JJ/VB/RB), followed by (ii) clause-based scoring, which computes clause-level valence means and selects the most extreme value.

purposes, temporal VAD lexicons thus serve as a measure of directional valence trajectories, rather than an absolute historical ground truth.

We thereby present temporally adjusted lexicons spanning 335,804 words, each assigned a real-valued score between -1 and 1 for the three VAD dimensions. A key challenge in this process lies in determining a suitable number of top-K neighbours. Selecting too few (top-10) can miss contextual diversity, but taking into account too many (top-500) may lead to semantic over-smoothing and can introduce noise. Following empirical analysis, averaging the scores from the top-20 neighbours seems to mitigate these adverse effects. To determine the number of neighbours required to retrieve 20 known words from the lexicon, we analysed the retrieval rates for 100 anchor words of high and low valence from Chronoberg. This analysis further revealed that retrieving 20 known neighbours typically necessitates inspecting the top-100 nearest neighbours (see Appendix A.4).

**Sentence-level VAD Annotations.** Extending our analysis beyond individual words, we leverage our five temporal VAD lexicons to assess sentiment in full sentences. However, a naive aggregation of word scores can lead to a neutrality bias, given the high frequency of neutrally perceived words in language. To mitigate this issue, we introduce two key elements in our scoring pipeline: JJ/VB/RB and clause-based averaging, illustrated in Figure 2.

*JJ/VB/RB averaging:* First, we focus only on emotionally salient parts of speech (in particular verbs (VB), adverbs (RB), and adjectives (JJ)), thereby minimizing the influence of neutral words on the final sentence score.

*Clause-based averaging:* Second, our method

accounts for sentiment variations within a sentence, following prior work (Wang et al., 2018). Accordingly, we calculate an average valence for each clause, based on its adjectives (JJ), verbs (VB), and adverbs (RB). The overall sentence score is then the minimum value among these clause-level scores.

*Final Score:* We compute the final score  $\hat{\mathbf{I}}_{\text{VAD}}(\text{sent})$  by combining these two approaches:

$$\min_{C_i \in \text{Clauses}} \left( \frac{\sum_{t \in C_i, \text{pos}(t) \in \{\text{JJ}, \text{VB}, \text{RB}\}} \mathbf{A}_{\text{VAD}}(t)}{N_{C_i, \{\text{JJ}, \text{VB}, \text{RB}\}}} \right), \quad (2)$$

where  $\mathbf{A}_{\text{VAD}}(t)$  represents the VAD scores and  $\text{pos}(t)$  is the part of speech tag for each token in the sentence. Since values range from -1 to +1, we will consider the sign to carry a respective connotation for simplicity in further analysis. This approach is particularly effective on sentences that combine positive and negative clauses, where naive averaging would otherwise mask the hateful content. We believe this to be justified by focusing on the assessment of the overall change in consecutive analysis. However, we acknowledge that perceived connotation can depend on various subjective factors. Our sentence-level valence annotations are a core component of Chronoberg and are publicly available to support transparency and further analysis.

## 2.4 Chronoberg Composition & Statistics

Finally, we highlight Chronoberg’s summary statistics. A dataset sheet for datasets (Gebu et al., 2021) is provided in Appendix C.

**Composition.** Overall, Chronoberg is composed of 2.7B tokens, representing 91M sentences from 25,061 English-language books published between 1750 and 2000, with additional metadata in the form of temporally-aligned VAD lexicons that span 335,804 words. On average, approximately 28% of sentences per 50-year epoch are classified as *negative* (valence  $< 0 - \epsilon$ ), while 50% are *positive* (valence  $> 0 + \epsilon$ ), considering  $\epsilon$  to be 0.05. Notably, some unique samples across epochs exhibit a change in average valence scores.

**Diachronic Shifts in Words & Sentences.** To distinguish semantically stable words from those undergoing genuine diachronic drift, we apply a threshold  $|\Delta| > 0.1$  to the endpoint difference between first and last available epoch valence estimates. A word is classified as *Shifting* if  $|v_{\text{last}} - v_{\text{first}}| > 0.1$  and the valence category

| Shift                   | Word          | Tier     | 1750s | 1800s | 1850s | 1900s | 1950s |
|-------------------------|---------------|----------|-------|-------|-------|-------|-------|
| <i>pos</i> → <i>neg</i> | asylum        | <b>H</b> | 0.27  | -0.24 | -0.54 | -0.52 | -0.65 |
|                         | weird         | <b>H</b> | 0.30  | 0.01  | -0.28 | -0.33 | -0.43 |
|                         | homeless      | <b>M</b> | 0.11  | -0.62 | -0.63 | -0.66 | -0.56 |
|                         | punk          | <b>M</b> | 0.20  | 0.14  | -0.25 | -0.17 | -0.26 |
|                         | erasure       | <b>L</b> | 0.10  | 0.02  | -0.06 | -0.12 | -0.13 |
|                         | astrologer    | <b>L</b> | 0.30  | 0.20  | 0.11  | 0.04  | 0.00  |
| <i>neg</i> → <i>pos</i> | febrile       | <b>H</b> | -0.58 | -0.53 | -0.66 | -0.54 | 0.33  |
|                         | infatuation   | <b>H</b> | -0.66 | -0.63 | -0.52 | -0.35 | 0.40  |
|                         | shack         | <b>M</b> | -0.67 | -0.05 | -0.02 | 0.30  | 0.10  |
|                         | managerial    | <b>M</b> | -0.65 | 0.25  | 0.20  | 0.26  | 0.30  |
|                         | heterogeneous | <b>L</b> | -0.21 | -0.17 | -0.13 | -0.03 | 0.07  |
|                         | modernism     | <b>L</b> | 0.08  | 0.14  | 0.20  | 0.20  | 0.30  |

Table 1: Examples of words with strong diachronic valence shifts in Chronoberg. Scores are averaged over top-20 nearest neighbours per epoch and range from -1 (maximally negative) to +1 (maximally positive). Confidence tiers: **H** = unambiguous, dominant VAD shift; **M** = partial/single-step shift; **L** = sub-threshold drift.

changes (positive  $\leftrightarrow$  negative  $\leftrightarrow$  neutral); otherwise it is classified as *Stable*. For the validation study in 3.1, we further assign a confidence tier to each evaluative word based on two thresholds applied jointly to local step-wise movement and net endpoint shifts: (i) **H**igh requires at least two consecutive deltas exceeding the noise floor and  $|\Delta_{\text{net}}| > 0.2$ : the VAD trajectory shows a clear, multi-step directional shift. (ii) **M**edium is assigned when local movement is present but the net shift is modest ( $|\Delta_{\text{net}}| \leq 0.2$ ), or when only a single significant step is observed: the VAD signal is detectable but not unambiguous. (iii) **L**ow covers two cases: words where no individual step exceeds the noise floor but sub-threshold drift accumulates directionally across epochs. Words with no significant movement on either measure are classified *Stable*. Across the full Chronoberg lexicon of 335,804 words, 44,049 are assigned **H**igh, 36,301 **M**edium, and 2,518 **L**ow, with the remainder classified *Stable*, indicating that roughly one quarter of the lexicon shows detectable affective drift over the 250-year window.

Table 1 presents representative examples illustrating all three tiers. Among *pos*→*neg* words, *asylum* (+0.27→ -0.65, **H**) pejorated as the word shifted from sanctuary to associations with mental institutions; *homeless* (+0.11→ -0.56, **M**) acquired negative valence through urbanisation, concentrated in a single epoch; *astrologer* (+0.30→ +

0.00, **L**) shows slow sub-threshold drift. Among *neg*→*pos* words, *febrile* (−0.58→ + 0.33, **H**) ameliorated as metaphorical usage overtook clinical sense; *managerial* (−0.65→ + 0.30, **M**) gained prestige; and *heterogeneous* (−0.21→ + 0.07, **L**) drifted slowly toward neutral as diversity became descriptive rather than evaluative.

These examples emphasize the value of our temporal VAD lexicons and Chronoberg as instruments for fine-grained analysis of semantic and affective change. Additional examples are in Appendix A.5.

### 3 Experiments and Dataset Analysis

We present three experiments demonstrating Chronoberg’s utility: validating the temporal VAD lexicons against words with documented semantic shifts, to assess if VAD trajectories capture such shifts; examining whether contemporary hate-speech detectors capture shifts in harmful language as seen in Chronoberg; and investigating the extent to which these systems rely on surface-level lexical cues rather than contextual understanding.

#### 3.1 External Validation of the VAD Lexicons

To assess whether Chronoberg’s temporal VAD scores capture documented shifts, we validate against 119 words with verified changes compiled from existing literature, covering both evaluative cases (pejoration, amelioration, contronym) and non-evaluative cases (metaphor, narrowing, generalization, subjectification, polysemy). Words whose shift falls outside Chronoberg’s measurement window are excluded as outliers ( $n=12$ ). Each word is assigned a confidence tier (**H**/**M**/**L**) using the criteria defined in 2.4, and each misrepresentation is assigned to one of three error categories: *direction mismatch* (‡) where the corpus-level trajectory runs counter to the documented direction of change, typically because a dominant literary meaning masks a shift that unfolded in spoken or informal registers; *register gap* (★), where the shifted meaning is under-represented in edited book text to produce measurable valence shift; and *bidirectional sense split* (†), where opposing senses coexist and cancel in the aggregate score.

Table 2(a) shows that 23 of 40 in-scope *evaluative* words are correctly represented, with agreement between observed valence trajectory and expected direction of change confirmed against expert annotations. Accuracy stratifies by tier: **H**-tier

| (a) Evaluative changes |            |           |           |        |
|------------------------|------------|-----------|-----------|--------|
| Type                   | Conf. Tier | Rep.      | Misrep.   | Reason |
| Pejoration (30)        | <b>H</b>   | 8         | 1         | ‡      |
|                        | <b>M</b>   | 7         | 9         | ‡      |
|                        | <b>L</b>   | 0         | 5         | ★      |
| Amelioration (9)       | <b>H</b>   | 4         | 1         | ‡      |
|                        | <b>M</b>   | 4         | 0         | –      |
| Contronym (1)          | <b>M</b>   | 0         | 1         | †      |
| <b>Total (40)</b>      |            | <b>23</b> | <b>17</b> |        |

| (b) Non-evaluative changes        |             |             |
|-----------------------------------|-------------|-------------|
| Type ( $n$ )                      | Val. stable | Val. shifts |
| Metaphor (21)                     | 5           | 16          |
| Narrowing (4)                     | 0           | 4           |
| Generalization (3)                | 1           | 2           |
| Subjectification (7)              | 1           | 6           |
| <i>Multiple coexisting senses</i> |             |             |
| Polysemy (5)                      | 0           | 5           |
| Domain / Technology (3)           | 1           | 2           |
| SemEval Stable (24)               | 9           | 15          |
| <b>Total (67)</b>                 | <b>17</b>   | <b>50</b>   |

Table 2: Validation of Chronoberg’s thresholded VAD lexicons against documented semantic shifts. In (a) Confidence tiers (Conf. Tier): **H**igh = unambiguous, dominant VAD shift in corpus; **M**edium = partial or single-step shift; **L**ow = sub-threshold drift invisible to VAD. In (b), *Val. stable* = valence correctly flat, *Val. shifts* = connotative drift driven by a shift in the distribution of coexisting word senses. **Note:** (‡): direction mismatch: corpus-level trajectory runs counter to expected direction (e.g. *bully*, *guy*). (★): register gap: shifted sense underrepresented in Chronoberg (e.g. *Fairy*, *Bent*) or obscured by a stable dominant sense (e.g. *Queen*). (†): bidirectional sense split: opposing senses coexist and cancel (e.g. *sanction*).

words are correct in 12 of 14 cases (86%), **M**-tier in 11 of 21 (52%), and **L**-tier in none, indicating that the lexicons are reliable when corpus signal is strong and that the weaker aggregate is driven by inherently ambiguous medium-confidence cases. Of the 9 **H**-tier pejorations, 8 are correctly represented (e.g., *awful*; *bastard*). **M**-tier captures an additional 7 pejorations (e.g., *bitch*, *dyke*, *prick*) where signal is present but weak or concentrated in a single epoch transition. For amelioration, 8 of 9 words are correctly represented: 4 at **H**-tier (e.g., *awesome*, *witch*) and 4 at **M**-tier (e.g., *minx*, *terrific*). The 17 misrepresentations include one complex case belonging to *contronyms* (*sanction*), whose bidirection split is understandably diluted in the VAD lexicons; the other 16 misrepresen-

| YEAR  | Sentences                                                                     | Hate-Check Models |        |      |         | Valence Score | Affective Sentiment |
|-------|-------------------------------------------------------------------------------|-------------------|--------|------|---------|---------------|---------------------|
|       |                                                                               | RoBERTa+Persp     | OpenAI | Qwen | Mistral |               |                     |
| 1750s | but I loathe you, you apache indian!                                          | 🚩                 | 🚩      | 🚩    | 🚩       | -0.50         | 👎                   |
|       | Where is the woman to strew the flowers?                                      | 🚩                 | ✓      | ✓    | ✓       | 0.14          | 👍                   |
|       | you horse-hair hypocrite, you!                                                | ✓                 | ✓      | ✓    | ✓       | -0.60         | 👎                   |
| 1800s | How I wish that you were black!—I detest your colour.                         | 🚩                 | 🚩      | 🚩    | 🚩       | -0.79         | 👎                   |
|       | The black parts of them must be cut away, when the skin is taken off.         | 🚩                 | ✓      | 🚩    | 🚩       | 0.02          | 👍                   |
|       | he redoubled his gayety and carelessness.                                     | ✓                 | ✓      | ✓    | ✓       | -0.71         | 👎                   |
| 1850s | The Irish are mostly very filthy and diseased.                                | 🚩                 | 🚩      | 🚩    | 🚩       | -0.66         | 👎                   |
|       | They carry faggots.                                                           | 🚩                 | 🚩      | 🚩    | 🚩       | 0.08          | 👍                   |
|       | An outlandish pair brought from New York for some tory hussy                  | ✓                 | ✓      | ✓    | 🚩       | -0.10         | 👎                   |
| 1900s | As a race, as a family, the blacks have no loyalty.                           | 🚩                 | 🚩      | 🚩    | 🚩       | -0.02         | 👎                   |
|       | I may cut you out of my gold expedition, if you get gay.                      | 🚩                 | 🚩      | 🚩    | 🚩       | 0.175         | 👍                   |
|       | We know how heartless he is, how vindictive, how horribly cruel.              | ✓                 | ✓      | 🚩    | ✓       | -0.77         | 👎                   |
| 1950s | I hope with the Negroes, Indians, .. reduce Virginia..                        | 🚩                 | 🚩      | 🚩    | 🚩       | -0.18         | 👎                   |
|       | Black should never be worn at a wedding.                                      | 🚩                 | 🚩      | ✓    | ✓       | 0.10          | 👍                   |
|       | Why are the Africans in Algeria rising against their white French oppressors? | ✓                 | ✓      | ✓    | ✓       | -0.05         | 👎                   |

Table 3: Sentence-level classifications in Chronoberg using LLM-based hate-check tools (RoBERTa+Perspective API, OpenAI Moderation API, Qwen and Mistral) (🚩 = Hate, ✓ = Non-hate) and valence-based scoring (👎 = Negative, 👍 = Positive sentiment). For each 50-year interval, we organize based on model agreement. The first row illustrates cases where all tools collectively classify an instance as harmful. Rows 2-3 show instances where they disagree, classifying them as either positive or negative. While VAD lexicons provide interpretable complementary signals, we acknowledge that harmful texts are inherently subjective; therefore, we do not regard them as definitive solutions to LLM misclassification, but rather as potential tools to enhance performance.

tations fall into two interpretable categories. 11 words show corpus-level movement in the *wrong direction* (‡): 9 medium-tier pejorations (e.g. *bully* and *egregious*) appear to ameliorate in book text because their dominant literary senses pull aggregate neighbourhood valence upward, masking the pejorative shift that occurred primarily in spoken and informal registers; the ameliorated word *guy* and the pejorated word *batty* similarly move in the opposite direction to expectation. Five words (*fatal*, *weasel*, *queen*, *bent*, *fairy*) are assigned **L**ow: their pejorated senses were too infrequent in edited book text to produce measurable valence movement, reflecting a register gap (★).

Table 2(b) depicts the result for non-evaluative changes, where no valence shift is expected. 8 words are correctly classified as valence-stable: 5 metaphor words (*attack*, *bug*, *edge*, *land*, *platform*) and 1 each from generalization (*head*), subjectification (*major*), and domain/technology (*computer*). The remaining 35 words not belonging to SemEval-2020 (Schlechtweg et al., 2020) that show valence movement are analytically informative rather than erroneous: non-evaluative changes such as technological metaphors (*network*, *film*) and subjectification (*actually*, *starting*) produce genuine connotative drift as a word’s dominant associative context shifts, pulling aggregate neighbourhood valence with it. Among the 24 SemEval words, 9 are cor-

rectly classified as stable and 15 show apparent valence movement. We note that SemEval-2020 defines stability over a short two-timepoint window, while Chronoberg spans 250 years; thus, some of these cases may reflect genuine long-horizon drift that falls outside SemEval’s detection threshold, rather than spurious correlation. Where misclassification does occur, it is attributable to representation gaps and polysemy dilution, which are expected limitations of a corpus-driven approach on historical text. These results also address a potential circularity in the propagation method 2.3, since VAD scores are propagated from contemporary NRC anchors, they could in principle reflect modern rather than historical associations. However, the reliable recovery of documented shifts precisely where corpus signal is strong (that is, **H**-tier cases) offers indirect evidence that such circularity does not dominate the propagated scores.

### 3.2 Do LLMs Capture Shifts in Harmful Language within Chronoberg?

Having validated that the lexicons capture documented semantic change, we next leverage our temporal VAD annotations to contextualize practical notions of harmful language over time. Hate speech represents particularly well-defined and inherently negatively connoted subsets, where temporal shifts in meaning are especially consequential. We em-

phasize that lexical valence is neither equivalent to nor a proxy for hate. A sentence may be hateful without strongly negative valence and vice versa. The propagated VAD scores serve throughout as interpretable diagnostic signals of temporal misalignment, not as ground truth for hate detection. Prior works have shown that hate-speech classifiers degrade under temporal shift, as token-level sensitivity causes models to flag words regardless of historical usage (Davidson et al., 2017). Cross-period generalisation also remains poor even with temporal lexical features (Röttger and Pierrehumbert, 2021; McGillivray et al., 2022), and static benchmarks actively mask both issues (Di Bonaventura et al., 2025). This allows us to investigate alignment between hate-check outputs and VAD scores, highlighting cases where our annotations capture meaningful affective polarity. More importantly, discrepancies in sentiment expose where modern classifiers fail to recognize historically situated expressions of hate or over-generalize from present-day keyword associations.

To identify suitability, we first evaluated nine hate-speech detection tools on HateCheck (Röttger et al., 2021) (see Appendix B). Based on this analysis, we adopt a two-stage pipeline: RoBERTa<sup>2</sup> (Vidgen et al., 2021) achieves the strongest recall ( $P=0.963$ ,  $R=0.973$ ) but is prone to false positives, while Perspective API (Lees et al., 2022) offers near-perfect precision at low recall ( $P=0.993$ ,  $R=0.389$ ); we therefore use RoBERTa to maximize coverage and Perspective API at a threshold of  $\geq 0.7$  to filter false positives. We also evaluate the OpenAI moderation API<sup>3</sup> as a recent baseline.

In Table 3, we show representative examples where VAD annotations and current hate-check tools agree and diverge. They are illustrative of notable trends in how modern hate-check tools flag negative sentences. While the latter achieve consistent correct predictions in cases of explicit sentiment (first row of each time interval), they seem to struggle when sentiment is implied rather than directly stated. For instance, a neutral phrase such as “*Black should never be worn at a wedding*” is incorrectly flagged as hateful by both hate-check tools, whereas the valence scores more accurately capture its neutral sentiment (0.10). Another illustrative case from the 1850s is the phrase “*tory*

<sup>2</sup>RoBERTa is a pretrained language model; we use “hate-check tools” as a collective shorthand.

<sup>3</sup><https://platform.openai.com/docs/guides/moderation>

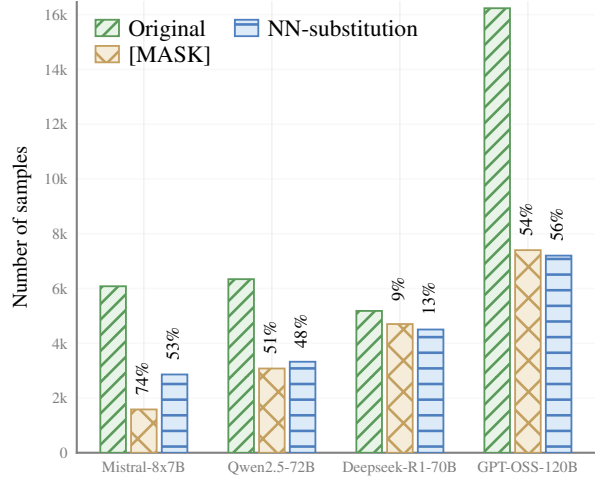


Figure 3: Substituting contemporary target words with historical equivalents improves the contextualization of modern LLMs for hate-speech identification. The green bars report the disagreement rates between LLMs and Chronoberg’s temporal VAD lexicons, while the orange and blue lines report the disagreement rates after applying two substitution strategies: replacing target words with a [MASK] token or a nearest-neighbour (NN) word from the corresponding era, across four models (Mistral, Qwen2.5, GPT, and DeepSeek). Both substitutions consistently reduce disagreement rates by 50-75% for three of our four models, highlighting the models’ over-reliance on surface-level lexical cues rather than historical context.

*hussy*”, where LLMs misinterpret *hussy* with its modern connotation, influencing the judgment of the sentence’s hatefulness.

### 3.3 Do Modern Hate-Check Tools Rely on Surface-Level Lexical Cues?

Discrepancies between LLM hate-speech classifiers and our temporal VAD lexicons reveal a systematic pattern of surface-level token sensitivity. For instance, LLM classifiers negatively connotate contexts such as “*They carry faggots*”, even though the term *faggot* commonly referred to bundles of wood and not the contemporary association with slurs (Michel et al., 2011). In contrast, our temporal VAD lexicons assign the sentence a positive valence score, reflecting its historically situated meaning. This leads to a hypothesis that classifier decisions are driven by the target token itself, rather than its sentential context. To validate this hypothesis at scale, we conduct systematic substitutions of words with period-appropriate synonyms in the corresponding VAD lexicon. For instance, the word *faggot* would be replaced with *firewood*, *sticks*, or *bundle* in an otherwise identical sentence. Specifi-

cally, we first identify all sentences in Chronoberg where RoBERTa and our VAD lexicons conflict, treating these disagreements as candidate cases of classifier overreach. We then screen these sentences across four contemporary LLMs (Mistral (Jiang et al., 2023), Qwen2.5 (Qwen Team et al., 2025), GPT-OSS (OpenAI, 2025), Deepseek (Guo et al., 2025)), applying a consistent substitution procedure: each flagged word is replaced with a historically accurate, semantically equivalent but non-toxic neighbour drawn from our Word2Vec embeddings, under two conditions: replacing target words with a [MASK] token or a nearest-neighbour (NN) word from the corresponding era. In Figure 3, we observe that a single substitution flips predictions from hateful to not-hateful for about 50–75% of the initially flagged sentences for three out of the four models. The results confirm that a simple lexical substitution is often sufficient to flip classifier predictions from hateful to not-hateful across all variants. Whereas the VAD score remains consistent throughout, modern classifiers remain sensitive to surface-level lexical cues and fail to account for diachronic shifts in word meaning. Chronoberg’s temporal VAD lexicons can thus serve as a complementary diagnostic mechanism for identifying these failures. Beyond diagnosis, they point towards temporally aware adaptation: since continual pretraining on target-period text recovers within-period but not cross-period performance (Röttger and Pierrehumbert, 2021), era-specific VAD trajectories could instead be provided to classifiers as an auxiliary signal indicating what a term meant and how it was perceived historically, directly targeting the surface-cue sensitivity demonstrated here.

## 4 Conclusion

We introduced Chronoberg, a large-scale, temporally structured corpus of English books spanning the years 1750-2000, enriched with diachronic VAD lexicons and sentence-level affective annotations. Using these resources, we quantified semantic and affective shifts, and demonstrated the need for modern LLM-based tools to better situate their detection of discriminatory language and contextualization of sentiment across various time periods. Chronoberg opens several promising avenues, such as broader continual- and lifelong-learning studies with temporal training pipelines, as well as machine unlearning protocols to address historically contingent slurs or outdated facts. Future work

includes strategic sampling of books containing homonyms or underrepresented words to update their meanings, exploring finer temporal granularities and drawing interdisciplinary connections to key historical eras or literary epochs.

## 5 Limitations

Although publication date inference introduces residual uncertainty (MAE  $\pm 3.05$ , SD 5.20 years), this is acceptable at the decadal granularity at which our analysis operates. We nonetheless recommend temporal bins no smaller than 15 years for future work. As retrospective human annotation at this scale is improbable, and expert linguistic validation across 335,804 words and five centuries equally impractical, the temporal VAD lexicons should be read as directional trajectory measures rather than absolute historical ground truth, with confidence tiers treated as indicative rather than definitive. Since Chronoberg’s literary corpus excludes the spoken and subcultural language where most pejorative shifts originate, failures on words such as *fairy* and *queen* are expected. Finally, although our substitution experiment demonstrates systematic surface-level sensitivity in LLM classifiers, hate speech is inherently subjective across historical periods and the binary hateful/non-hateful framing remains too coarse for nuanced historical analysis. Thus, the lexicons should accordingly be treated as a diagnostic tool in conjunction with moderation tools rather than as a replacement.

## 6 Ethical Considerations

All texts in Chronoberg are sourced from (Project Gutenberg)’s public domain collection, including some that reference identifiable historical figures and describe subpopulations by race, gender and religion; we caution against using the dataset to single out individuals or groups. The dataset contains historically situated language that readers may find offensive, including slurs and discriminatory expressions reproduced solely for scientific analysis. As the corpus reflects the literary landscape from the perspective of prolific authors between 1750 and 2000, findings should not be generalized beyond this time frame. The hate-speech annotations included in Chronoberg are a diagnostic resource; they should not be used as training data for modern classifiers without careful mitigation, as doing so risks encoding historical prejudices into them.

## References

- Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte Im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*.
- Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. A deep dive into multilingual hate speech classification. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2020), Applied Data Science and Demo Track*.
- Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- badmatr11x. 2023. *DistilRoBERTa base offensive and hateful speech text multiclassification*. Hugging Face. Hugging Face model repository.
- Naomi Baes, Nick Haslam, and Ekaterina Vylomova. 2024. A multidimensional framework for evaluating lexical semantic change with social science applications. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Andreas Blank. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*. Max Niemeyer Verlag.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Leonard Bloomfield. 1933. *Language*. Holt, Rinehart and Winston, New York.
- Michel Bréal. 1897. *Essai de sémantique: science des significations*. Hachette, Paris.
- John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*.
- Jáder Martins Camboim de Sá, Marcos Da Silveira, and Cédric Pruski. 2024. Survey in characterization of semantic change. *arXiv preprint arXiv:2402.19088*.
- Pierluigi Cassotti, Naomi Baes, Stefano De Pascale, Jáder Martins Camboim de Sá, Francesco Periti, Nick Haslam, Dirk Geeraerts, and Nina Tahmasebi. 2026. SenseRel: A sense-level benchmark for denotational and connotational meaning relations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Pierluigi Cassotti, Stefano De Pascale, and Nina Tahmasebi. 2024. Using synchronic definitions and semantic relations to classify semantic change types. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the Conference on Hypertext and Social Media*.
- Thomas Davidson, Dana Warmesley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*.
- Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English (COHA). *Corpora*, pages 121–157.
- Mark Davies. 2015. Corpus of contemporary American English (COCA). <https://doi.org/10.7910/DV/N/AMUDUW>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.
- Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, pages 257–273.
- Chiara Di Bonaventura, Barbara McGillivray, Yulan He, and Albert Meroño-Peñuela. 2025. Hatevolution: What static benchmarks don’t tell us. In *Findings of the Association for Computational Linguistics: ACL 2025*.
- Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training temporal word embeddings with a compass. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Daniel Edmondson. 2022. Word norms and measures of linguistic reclamation for LGBTQ+ slurs. *Pragmatics and Cognition*.

- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Communications of the ACM*.
- Google Books. 2026. <https://developers.google.com/books>. Accessed:2026-01-19.
- Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*. Association for Computational Linguistics.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and DeepSeek-AI. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*.
- 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*.
- Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*. IEEE.
- Robert J. Jeffers and Ilse Lehisté. 1979. *Principles and Methods for Historical Linguistics*. MIT Press.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Christian Kay, Jane Roberts, Michael Samuels, and Irené Wotherspoon. 2009. *Historical Thesaurus of the Oxford English Dictionary*. Oxford University Press, Oxford.
- Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Petra Kralj Novak, Teresa Scantamburlo, Andraž Pelićon, Matteo Cinelli, Igor Mozetič, and Fabiana Zollo. 2022. Handling disagreement in hate speech modelling. In *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2022)*.
- Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*.
- Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liška, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d’Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*.
- Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective API: Efficient multi-lingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’22)*. Association for Computing Machinery.
- Adrienne Lehrer. 1985. Markedness and antonymy. *Journal of linguistics*, pages 397–429.
- Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.
- Jeffrey Li, Mohammadreza Armandpour, Seyed Iman Mirzadeh, Sachin Mehta, Vaishaal Shankar, Raviteja Vemulapalli, Samy Bengio, Oncel Tuzel, Mehrdad Farajtabar, and Hadi Pouransari. 2025. TiC-LM: A web-scale benchmark for time-continual LLM pre-training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Library of Congress Linked Data Service. 2026. <https://id.loc.gov/>. Accessed:2026-01-19.
- Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! Analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

- Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- 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 First Workshop on Ever Evolving NLP (EvoNLP)*.
- Manisha Mehta and Fausto Giunchiglia. 2025. Understanding Gen Alpha’s digital language: Evaluation of LLM safety systems for content moderation. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3.
- Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014).
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Saif Mohammad. 2018. Obtaining reliable human ratings of Valence, Arousal, and Dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.
- Saif M Mohammad. 2025. NRC VAD Lexicon v2: Norms for Valence, Arousal, and Dominance for over 55k English terms. *arXiv preprint arXiv:2503.23547*.
- Kohei Oda, Hiroya Takamura, Kiyoaki Shirai, and Natthawut Kertkeidkachorn. 2025. Improving interpretability of lexical semantic change with neurobiological features. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*.
- OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.
- OpenLibrary. 2026. <https://openlibrary.org/developers/api>. Accessed:2026-01-19.
- Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. 1957. *The measurement of meaning*. University of Illinois Press.
- Matjaž Perc. 2012. Evolution of the most common English words and phrases over the centuries. *Journal of The Royal Society Interface*.
- Juan Manuel Pérez, Juan Carlos Giudici, and Franco M. Luque. 2021. pysentimiento: A python toolkit for opinion mining and SocialNLP tasks. *arXiv preprint arXiv:2106.09462*.
- Project Gutenberg. 2025. <https://www.gutenberg.org>. Accessed: 2025-12-01.
- ProQuest. 2008. Early English Books Online (EEBO). <https://proquest.libguides.com/eebopqp>.
- Organizers Of Queerinaï, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, and 31 others. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.
- Paul Röttger and Janet B. Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.
- James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological Review*, pages 145–172.
- 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*. International Committee for Computational Linguistics.
- Text Creation Partnership. 2008. Early English Books Online Text Creation Partnership (EEBO-TCP). <https://www.textpartnership.net/pages/eebo-tcp.html>.
- Elizabeth Closs Traugott. 2017. Semantic change. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Elizabeth Closs Traugott and Richard B. Dasher. 2001. *Regularity in Semantic Change*. Cambridge University Press.

Stephen Ullmann. 1962. *Semantics: An Introduction to the Science of Meaning*. Basil Blackwell, Oxford.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.

Jingjing Wang, Jie Li, Shoushan Li, Yangyang Kang, Min Zhang, Luo Si, and Guodong Zhou. 2018. Aspect sentiment classification with both word-level and clause-level attention networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*.

Wikipedia. 2026. [https://www.mediawiki.org/wiki/API:Main\\_page](https://www.mediawiki.org/wiki/API:Main_page). Accessed:2026-01-19.

## Appendix

We have organized our supplementary material in the following order:

- **Section A: Dataset Curation and Lexical Analysis.** Starting with an extended related works section, we then provide additional details on dataset collection, curation, and filtering pipeline, along with an extended analysis of VAD lexicons.
- **Section B: Extended Hate-speech Analysis.** Expanded evaluation of harmful texts in Chronoberg, including comparisons across multiple hate-speech detection tools, sentence-level analysis, and highlighting cases of disagreement between different models and valence scores in hate prediction.
- **Section C: Datasheet for Chronoberg.**

### A Additional Chronoberg dataset details

This appendix complements main body Section 2 by providing additional information on the choice of data sources, available metadata, and recovery of publication dates through inference methods. Each of the following subsections expands upon specific components of the dataset pipeline to clarify design decisions and practical challenges. We also provide insights into the distribution and the thematic composition of Chronoberg.

#### A.1 Extended Related Works

**Diachronic corpora and temporal datasets.** The earliest large-scale temporal corpora such as Google Books (Michel et al., 2011) and Early English Books Online (EEBO) (ProQuest, 2008; Text Creation Partnership, 2008) laid out the groundwork for analyzing quantitative phenomena at the interface of social sciences and humanities. However, the lack of sentence context and semantic annotations at an n-gram level preclude sentence-level inference and direct use in timely LLM fine-tuning or evaluation. Other diachronic corpora such as COHA (Davies, 2012), COCA (Davies, 2015), and CCOHA (Alatrash et al., 2020), are valuable for capturing American English variation, but remain too small in scale to train, so as to meaningfully probe modern-scale models. Newer resources like TiC-LM (Li et al., 2025) and TemporalWiki (Jang et al., 2022) emphasize contemporary factual

content rather than gradual connotative and affective evolution that characterises long-term semantic change. Chronoberg complements these prior efforts by providing a literary corpus containing full-length texts with temporal metadata, facilitating both semantic and affective analysis at yearly granularity in the context of modern-day LLMs.

#### Distributional semantics and VAD lexicons.

The distributional hypothesis (Bullinaria and Levy, 2007) forms the basis of all modern approaches to semantic change, where meaning is approximated by context and shifts in context distribution signals shifts in meaning. Count based methods using SVD (Levy et al., 2015) were superseded by predict-based embeddings in Word2vec (Mikolov et al., 2013), whose temporal extension through alignment techniques like Compass-aligned distributional embeddings (CADE) (Di Carlo et al., 2019) enables temporal comparisons without the instability of naive concatenation. Validation studies confirm that these methods recover documented changes, through cosine displacement (Hamilton et al., 2016), change-point detection (Kulkarni et al., 2015), community annotation in SemEval-2020 (Schlechtweg et al., 2020) as well as typological approaches (Blank, 1997; Kay et al., 2009; Cassotti et al., 2024). Critically, however, these benchmarks operate on small curated word sets, which makes them unsuitable for evaluating downstream behaviour at scale.

The affective dimension adds a further layer of complexity. Rooted in influential factor analysis (Osgood et al., 1957; Russell, 2003), the VAD framework decomposes word meaning into three dimensions: *valence* (the positive/negative evaluative quality of a word), *arousal* (the active/passive or exciting/calm tone) and *dominance* (the dominant/submissive or degree of control or power the word conveys). These dimensions have been human-annotated for 45,000 contemporary English words (Mohammad, 2018, 2025), providing fine-grained affective grounding. Diachronic embedding studies further demonstrate that such affective associations shift systematically over decades (Garg et al., 2018), reflecting broader cultural and social change. Recent works incorporate affective dimensions directly within semantic change frameworks, through multidimensional evaluation schemes that connect lexical change to social science constructs such as sentiment and intensity (Baes et al., 2024) and sense-level benchmarks

for denotational and connotational meaning relations (Cassotti et al., 2026). Surveys examine how change types such as pejoration and amelioration map onto connotational dimensions (Camboim de Sá et al., 2024), while other works link semantic change to interpretable affective features (Oda et al., 2025). Similarly, computational analysis of dehumanizing language connects affective lexical representations to social evaluation (Mendelsohn et al., 2020). As with semantic change benchmarks above, these approaches operate on curated word sets rather than corpus-scale lexicons.

Yet existing VAD lexicons are synchronic in nature; the scores reflect only contemporary linguistic understanding and do not account for historical semantic evolution. Consequently, the resource cannot track the changing connotations of words such as *broadcast* (originally denoting the physical scattering of seeds before acquiring its modern media sense), or *febrile* (which once carried strong connotations of danger and mortality before becoming a relatively neutral clinical descriptor), as modern ratings fail to capture their historical usage (Perc, 2012). Chronoberg addresses this gap by contributing computationally constructed, temporally aligned VAD lexicons spanning 335,804 words across five 50-year epochs between the years 1750-2000, which are then used for sentence-level VAD annotations across the full corpus, enabling the study of affective change and temporal robustness at the scale required for LLM safety evaluation.

**Temporal robustness of language models.** The temporal degradation of LLMs is well-documented at the task level, with works showing that perplexity rises monotonically with temporal distance from the training window (Lazaridou et al., 2021) and also how this generalization gap persists across several NLP tasks despite domain-adaptive pretraining (Luu et al., 2022). A key insight from both the studies is that the gap is not merely a knowledge problem, where models simply lack recent facts, but rather a distributional one. Surface statistics, syntax and word usage all shift in ways that trained models cannot anticipate. Works on temporal adaptation (Röttger and Pierrehumbert, 2021) show that continual pretraining on target-period text recovers performance within that period but does not improve cross-period generalization, suggesting that temporal robustness requires evaluation resources that span long horizons rather than adjacent time

windows. This has also been framed as a training pipeline problem (Dhingra et al., 2022), which argues for time-stamped corpora such as Chronoberg as a prerequisite for temporally aware models.

**Hate-speech detection and temporal bias.** The surface-token problem in hate-speech detection, where classifiers are triggered by the presence of a slur regardless of syntactic role or pragmatic context, was identified in (Davidson et al., 2017), where the lexical classifier struggled to classify Tweets and sentences without explicit hate keywords. This conflation of token identity with intent is precisely the failure Chronoberg has been designed to quantify. The temporal dimension compounds the problem, with works showing that classifiers trained on social-media text from one period generalize poorly even to the following year (Röttger and Pierrehumbert, 2021), while others demonstrate that incorporating temporal lexical features can yield measurable improvements (McGillivray et al., 2022). A more comprehensive study (Di Bonaventura et al., 2025) evaluated several language models and found that static test sets underestimate the effect of temporality; models that appear robust on fixed benchmarks fail when data drawn from a shifted distribution was used. Broader surveys on fairness and bias in hate-speech detection (Blodgett et al., 2020; Röttger et al., 2021) further establish this finding. Chronoberg contributes computationally constructed temporally aligned VAD lexicons, which are then used for sentence-level VAD annotations across the dataset, making the study of the interaction between connotative drift and classifier sensitivity empirically tractable.

## A.2 Data Source

As outlined in our dataset construction pipeline, we now present a detailed explanation of the rationale behind our choice of data corpus. The initial step in constructing a text dataset involves identifying an appropriate data source. For temporal datasets, two primary criteria are essential: first, the availability of timestamps indicating when the data was created; second, a substantial volume of content to ensure comprehensive coverage across different time periods. While large datasets are generally beneficial for language modelling since their performance improves with the amount of training data, temporal datasets require extensive data to capture variations over time.

In order to meet these requirements, we seek extensive text collections produced over the past centuries, easily accessible for research purposes, and accompanied by metadata detailing date of creation. Books emerged as a natural solution since they offer coherent and curated content, especially when compared to shorter form content like news or social media posts.

However, using books for large-scale, temporally annotated datasets presents several practical challenges:

1. **Copyright Restrictions:** Many books are under copyright restrictions, limiting free access to their full texts.
2. **Digitization Requirements:** To be usable, books must be available in digital formats.
3. **Metadata Availability:** Metadata, such as accurate author names and publication years, is crucial, since manually annotating hundreds of thousands of books without this information is unfeasible.
4. **Programmatic Access:** Efficient data collection necessitates programmatic interaction with the data source, such as through APIs, to download and filter relevant books in bulk.

We examined various online book databases, including Google Books ([Michel et al., 2011](#)), the Internet Archive, and Project Gutenberg ([Project Gutenberg](#)), to assess their suitability for large-scale historical text collection. Google Books offers an extensive online library, with full-text search across a vast collection of books and metadata such as publication dates. However, the API imposes significant constraints, limiting queries to a maximum of 40 results per request. Pagination requires numerous inefficient calls, and large-scale automated retrieval is hindered by protective measures such as CAPTCHAs. Moreover, the API requires a keyword-based search and does not permit queries based solely on publication year, making systematic dataset construction difficult.

The Internet Archive provides a large repository of digitized texts, often with richer metadata than Google Books. However, the quality and completeness of metadata varies considerably across the entries, and publication year information is frequently missing or unreliable. Additionally, bulk access is limited by rate restrictions and heterogeneity of

formats, which further complicates large-scale pre-processing.

Project Gutenberg is a widely used digital library of public domain literary works, providing unrestricted access to full-length texts. Each entry is accompanied by metadata covering attributes such as *title*, *authors*, *subjects* and *issue date*, as shown in Table 4, together with other metadata fields available. We therefore select this as the primary resource for curating our Chronoberg dataset. However, key metadata, most notably publication dates, are frequently absent or inconsistent, which poses challenges for constructing a coherent, temporally stratified dataset. The methodological details for our curation process to address this are discussed in the following section.

### A.3 Dataset Collection and Curation

Project Gutenberg contains some metadata inaccuracies, most notably the original year of publication often reflects the date of digitization rather than the actual release year. Accurate publication dates are critical for curating Chronoberg, as our aim is to order books chronologically for further study.

**Metadata and External Sources.** To address this limitation, we leverage available metadata attributes such as title, author, and the author’s birth and death dates. We also queried multiple external sources such as Google Books ([Google Books](#)), Google Search, the Library of Congress API ([Library of Congress Linked Data Service](#)), Open Library ([OpenLibrary](#)), Wikipedia ([Wikipedia](#)), using book titles and author lifespans to determine the correct publication year. While dates explicitly mentioned in book titles can provide additional cues when available, this method is applicable only to a limited subset of works. To assess the reliability of these tools, we manually annotated 100 books from Project Gutenberg, evenly spread between 1611 and 1912. Due to access limitations, Google Books and Google Search were excluded, and the Library of Congress API was discarded because of poor performance.

Among the two remaining methods, shown in Table 5, Open Library was the most reliable, correctly estimating publication years for 49 out of 100 tested books, with a Mean Average Error (MAE) of 3.05 years and an accuracy (acc@5, i.e. within 5 years margin) of 76%, which falls within the acceptable tolerance range for our analysis. The Wikipedia API retrieved years for

| Attribute                        | Explanation                                                                                                                                | In Catalogue |
|----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|--------------|
| <b>ID</b>                        | A real number assigned by Project Gutenberg to uniquely identify the eBook                                                                 | ✓            |
| <b>Type</b>                      | Text (>98 %), dataset, sound, image [...]                                                                                                  | ✓            |
| <b>Issued</b>                    | Release date of the book                                                                                                                   | ✓            |
| <b>Title</b>                     | The title of the book                                                                                                                      | ✓            |
| <b>Language</b>                  | The language in which the book is available                                                                                                | ✓            |
| <b>Authors</b>                   | All authors of the eBook                                                                                                                   | ✓            |
| <b>Subjects</b>                  | Library of Congress subject headings                                                                                                       | ✓            |
| <b>LoCC</b>                      | Library of Congress entries                                                                                                                | ✓            |
| <b>Bookshelves</b>               | Hand-curated eBook collections supplemented by 64 “Browsing” categories which were automatically assigned to mimic browsing in a bookstore | ✓            |
| <b>Publisher</b>                 | The publisher of the book                                                                                                                  | ✗            |
| <b>License &amp; Rights</b>      | Specifies the book’s copyright status (e.g public domain in the USA) and, when applicable, the specific license governing its use.         | ✗            |
| <b>Downloads</b>                 | How often the book has been downloaded                                                                                                     | ✗            |
| <b>Birth/death dates</b>         | Birth and death rates of all authors and translators if available                                                                          | ✗            |
| <b>Description &amp; MARC520</b> | Description and Summary of the eBook                                                                                                       | ✗            |
| <b>Translators</b>               | All translators of the eBook (if any)                                                                                                      | ✗            |
| <b>Datatypes</b>                 | E.g. text, HTML, ePub, PDFs, [...]                                                                                                         | ✗            |

Table 4: Project Gutenberg metadata fields with descriptions and catalogue availability. While many bibliographic attributes are present (✓), crucial information such as original publication dates is missing (✗), necessitating external inference for the construction of Chronoberg.

25 books with an accuracy of 42.0%. However, due to the limited sample sizes for Wikipedia, its reliability remains uncertain. The Open Library API provided sufficiently accurate data to support the construction of a chronological text corpus, particularly because we constrained errors within the author’s lifespan.

**Filtering.** The publication dates were constrained to fall within the author’s year of birth and death as an additional consistency check. In case of missing information, a default lifespan of 100 years was assumed. When API methods returned multiple year estimates, the most frequently occurring year (i.e., the modal year) was selected; in the case of a tie, the first appearing year was chosen. Further filtration criteria were applied to refine Chronoberg:

- (i) *Publication Year:* Books lacking any inferrable publication year (40.7%) were excluded. We restricted the dataset to books

| Scores                     | Open Library | Wikipedia | Majority Vote |
|----------------------------|--------------|-----------|---------------|
| Values                     | 92           | 47        | 96            |
| Correct                    | 49           | 25        | 41            |
| Accuracy @0 (%)            | 53.3         | 53.2      | 42.7          |
| Accuracy @5 (%)            | 76.0         | 39.0      | 74.0          |
| Accuracy @7 (%)            | 79.0         | 42.0      | 80.0          |
| MAE (years)                | 3.05         | 3.36      | 4.05          |
| Standard Deviation (years) | 5.20         | 7.24      | 6.56          |

Table 5: Comparison of publication year inference methods against 100 manually annotated ground truth samples. Accuracy is reported at different tolerances, with acc@0 representing the exact year, acc@5 within 5 years and acc@7 within 7 years. Open Library emerges as the most reliable predictor with broad coverage and low error (MAE = 3.05 years), making it the most reliable stand-alone source. Wikipedia achieves limited recall and lower accuracy. Majority voting across predictors offers marginal recall gains but does not scale well, reinforcing our choice of Open Library as the default predictor.

published between 1750 and 2000. This range balances token availability per decade and ensures linguistic consistency, as early modern English had largely been superseded by late

modern English by 1750. Additionally, this timeframe allows for the analysis of historical shifts in public perception and hate-speech evolution by contrasting older texts with those from the modern era.

- (ii) *Author Metadata*: Only books with a known author and recorded birth year were retained, leading to the exclusion of 22.9% of volumes. To ensure plausibility, we retained only works published within the author’s lifespan; for authors with unknown death years, this corresponds to a default cut-off of 100 years after birth. This step removes posthumous editions and maintains consistency.
- (iii) *Language*: The dataset was limited to English-language texts (80.3%) of the total books to reduce cross-linguistic variation and maintain consistency in temporal language analysis.
- (iv) *Translations*: Translated works (8.4%) were removed, since their publication dates often deviate substantially from the original text, potentially distorting historical trends.
- (v) *Content Type*: Non-textual materials such as images and audio files (1.7%) were excluded to preserve a purely text-based corpus.
- (vi) *Copyright and Availability*: Only books explicitly marked as public domain in the U.S. (98.8%) were included. Of these, 75 books lacked downloadable plaintext files in Project Gutenberg.

With these filtering steps, we successfully annotated 25,061 for chronological sorting out of the 73,500 books available in Project Gutenberg.

**Topic Distribution.** The distribution of books and tokens in Chronoberg across decades is uneven, as illustrated in Figure 4, reflecting the availability of texts in Project Gutenberg. The number of books increases steadily up to 1920s, a trend likely fuelled by population growth, educational expansion, and economic development. A sharp decline follows, primarily due to copyright restrictions, with most texts published post-1929 remaining under copyright, and consequently being unavailable in Project Gutenberg. As a result, early 20th-century works are strongly represented, while the post-1920 are under-represented. This is especially evident for the 2000s, which contain only the year 2000,

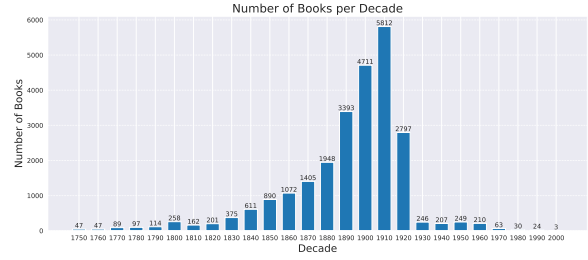


Figure 4: Temporal distribution of books in Chronoberg across decades (1750 - 2000). The number of available texts grows steadily until the 1920s, likely reflecting improved literacy and publication, but drops sharply thereafter due to copyright restrictions on works published after 1920. Consequently, early 20th century works dominate the distribution, while post-1920s coverage remains sparse.

| <i>Subjects</i>         |       |      |
|-------------------------|-------|------|
| Subject                 | Books | %    |
| Fiction                 | 5998  | 23.9 |
| History                 | 3012  | 12.0 |
| Juvenile fiction        | 2368  | 9.4  |
| Social Life and Customs | 1450  | 5.8  |
| 19th century            | 1218  | 4.9  |
| England                 | 1154  | 4.6  |
| United States           | 1052  | 4.2  |
| Great Britain           | 872   | 3.5  |
| Description and Travel  | 830   | 3.3  |
| Conduct of Life         | 695   | 2.8  |

| <i>Bookshelves</i>             |       |     |
|--------------------------------|-------|-----|
| Bookshelf                      | Books | %   |
| American Bestsellers 1895–1923 | 308   | 7.2 |
| Science Fiction                | 285   | 6.7 |
| Children’s Fiction 1895–1923   | 282   | 6.6 |
| Children’s Series 1895–1923    | 207   | 4.9 |
| Children’s Literature          | 203   | 4.8 |
| World War I                    | 196   | 4.6 |
| US Civil War                   | 192   | 4.5 |
| Historical Fiction             | 186   | 4.4 |
| Humor                          | 100   | 2.4 |
| Native America                 | 98    | 2.3 |

Table 6: Top 10 most frequent subjects and bookshelf topics in Chronoberg. Proportion of books assigned to each category are shown in % relative to the total number of books with at least one entry in the same category. “Subjects” are derived from Library of Congress headings, while “Bookshelves” are mostly from automatically assigned “browsing” categories. Note that multiple subject headings and bookshelves may be assigned to a single book.

producing a notably low count of books and tokens for that decade.

Beyond temporal coverage, Project Gutenberg metadata also provides insights into Chronoberg’s thematic composition. Table 6 presents the ten most common subjects and bookshelf topics, re-

| 1750–1799            |       | 1800–1849              |       |
|----------------------|-------|------------------------|-------|
| Author               | Books | Author                 | Books |
| Gibbon, Edward       | 14    | Bulwer-Lytton, Edward  | 88    |
| Schiller, Friedrich  | 7     | Scott, Walter          | 44    |
| Paine, Thomas        | 7     | Marryat, Frederick     | 40    |
| Wollstonecraft, Mary | 7     | Dickens, Charles       | 39    |
| Stanhope, Philip     | 7     | Cooper, James Fenimore | 33    |

| 1850–1899                  |       | 1900–1949                  |       |
|----------------------------|-------|----------------------------|-------|
| Author                     | Books | Author                     | Books |
| Twain, Mark                | 104   | Baum, Lyman Frank          | 73    |
| Ballantyne, Robert Michael | 88    | Stratemeyer, Edward        | 59    |
| Henty, George Alfred       | 74    | Barbour, Ralph Henry       | 57    |
| Alger, Horatio Jr.         | 72    | Oppenheim, Edward Phillips | 55    |
| Fenn, George Manville      | 66    | Wells, Carolyn             | 54    |

| 1950–2000                |       | All time                   |       |
|--------------------------|-------|----------------------------|-------|
| Author                   | Books | Author                     | Books |
| Leinster, Murray         | 17    | Twain, Mark                | 118   |
| Duellman, William Edward | 11    | Bulwer-Lytton, Edward      | 104   |
| Dick, Philip K.          | 10    | Ballantyne, Robert Michael | 88    |
| Kjelgaard, Jim           | 10    | Fenn, George Manville      | 87    |
| Norton, Andre            | 10    | Henty, George Alfred       | 86    |

Table 7: Chronoberg contains more than 20,000 authors, out of which the most represented ones are shown and grouped by historical period (1750–2000). Book counts indicate the number of works attributed to each author within the dataset and illustrate shifts in author prominence over time.

flecting the historical context in which the books were written. Fiction, in its various forms, is the most prominent genre in Chronoberg. Many works also reflect social and historical contexts followed by works addressing social issues and major historical events such as World War I and U.S. Civil War. Historical works covering earlier periods are also well-represented. In contrast, children’s literature constitutes a small portion of the dataset: the three bookshelf categories related to children’s literature account for only 692 books, forming a relatively niche subject.

The subjects used to categorize the books are drawn from the Library of Congress Subject Headings, whereas the categories for bookshelves are derived from a mixture of hand-curated eBook collections and automatically assigned “browsing” cat-

egories. Multiple subject headings or bookshelves can be assigned to a single book. The completeness of the metadata varies, where only 22 books in Chronoberg lack subject headings, but around 83% not having any bookshelf assigned. These were excluded from the counts shown in Table 6.

We also examine the authors represented in Chronoberg over time. Out of the 20k authors in Chronoberg, the most represented authors are shown in Table 7, illustrating shifts in author prominence across 250 years. While several prolific writers (e.g., Dickens, Schiller, Twain) appear prominently, the corpus is largely shaped by authors who were popular in their time, but are rarely read today. This distribution reflects typical biases of large-scale digitized historical corpora, capturing the broader landscape of historical print culture, in-

cluding many commercially popular authors whose work may better represent everyday linguistic usage of their period.

#### A.4 Investigating the effect of the number of neighbours on VAD measures

As discussed in Section 2.3 of the main text, the number of neighbours used to compute lexical scores is a crucial hyperparameter when constructing temporally aligned VAD lexicons. This choice directly affects the resulting VAD scores, where too few neighbours can introduce strong biases, while too many can produce overly neutral scores. Figure 5 illustrates this effect across the different eras for a select set of examples.

From Figure 5 it is observable that using the top-20 neighbours is typically adequate. However, because Chronoberg’s lexicons are based on the contemporary NRC VAD lexicon, not all top-20 neighbours in historical sentences may have corresponding NRC VAD scores. To address this, we conducted a small experiment, where we randomly sampled 100 words and measured how many nearest neighbours must be traversed to identify the 20 words present in the NRC VAD lexicon. The results in Figure 6 show that for most words, considering the top 50 neighbours is sufficient to obtain 20 valid scores, and the top 100 neighbours provide a reliable upper bound covering all words. Based on this analysis, we adopt the top-100 neighbours as the standard for computing VAD scores in Chronoberg, ensuring robust and consistent affective annotations.

#### A.5 Further examples for semantic shifts of words in Chronoberg

Figure 7 reports the distribution of valence shifts between the selected epochs. We present more qualitative examples that illustrate the transition of affective connotations, both from positive to negative and vice versa in Table 8. For instance, “sanctimonious”, which once conveyed genuine holiness, has shifted to denote a hypocritical display of moral superiority. Likewise, weird, originally associated with the supernatural and unearthly, now predominantly means “odd,” “strange,” or “bizarre.” Another compelling case is “depressive”, which evolved from a neutral or even positive association of “pressing down” to a word carrying strongly negative emotional connotations.

Conversely, we also observe cases of semantic shift from negative to positive. Words such as in-

fatuation, destiny, tweak, and repertoire exemplify this trend. “Infatuation” has moved from its earlier sense of “making foolish” to its modern meaning of intense admiration. Similarly, “tweak”, once meaning “to pluck or pinch,” has broadened to signify the act of making small adjustments. We have also reported several instances of words that were predominantly positive or negative across the time interval of 250 years, as shown in Table 9.

#### A.6 External Validation of Temporal VAD lexicons

**Method.** To assess whether Chronoberg’s temporal VAD lexicons capture documented semantic shifts, we apply a two-threshold confidence heuristic to each word’s valence trajectory across five 50-year epochs. Words with fewer than three observed values are excluded. For each remaining word, we compute consecutive per-epoch deltas between adjacent observed values, skipping missing epochs without interpolation. A step is *significant* if  $|\Delta_i| > 0.1$ . We also compute the overall net shift  $\Delta_{\text{net}} = v_{\text{last}} - v_{\text{first}}$  independently. Confidence tiers are assigned as follows: **H**igh requires at least two significant steps *and*  $|\Delta_{\text{net}}| > 0.2$ ; the trajectory shows a clear, multi-step directional shift. **M**edium is assigned when local movement is present but the net shift is weak ( $|\Delta_{\text{net}}| \leq 0.2$ ), or only a single significant step is observed: signal is detectable but not unambiguous. Finally, **L**ow is assigned when the shift is invisible to the VAD lexicons, either because no step exceeds the noise floor or because only sub-threshold drift accumulates. *Stable* is assigned when no meaningful signal is present on either measure. Direction is assigned following this procedure: if  $|\Delta_{\text{net}}| > 0.2$ , direction follows the sign of the net shift directly; if a linear regression over the observed epoch values yields  $R^2 \geq 0.3$  and  $|\Delta_{\text{net}}| > 0.1$ , direction follows the sign of the regression slope; otherwise direction is determined by the dominant sign among significant consecutive steps, with net shift sign as tiebreaker. Table 2 in the main paper reports aggregated results; full per-word trajectories appear in Tables 11–15 below.

**Evaluative Changes.** Of the 40 in-scope evaluative words, 23 are correctly detected by the thresholded VAD lexicons. 8 **H**-tier pejorations agree with the expected direction, confirming that clear, dominant evaluative shifts produce reliable VAD signals. Representative

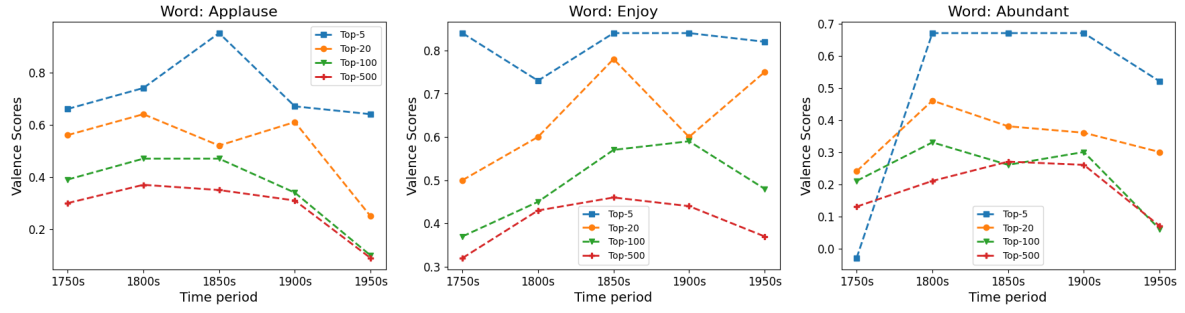


Figure 5: Visualizing the variance in the valence scores across the different time intervals. We vary the number of the Top-K neighbours to compute the affective valence scores for each word. Top-5 neighbours lead to strong bias, whereas top-500 neighbours lead to neutrality.

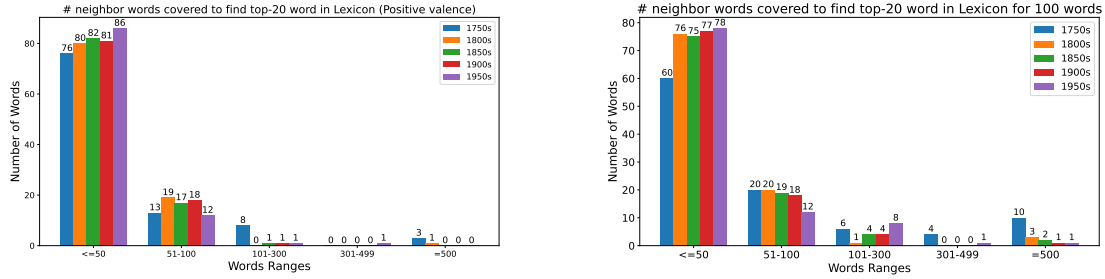


Figure 6: Visualizing the number of neighbours needs to be traversed to determine a VAD score by identifying the top-20 words in NRC VAD Lexicons. For most words, top-50 neighbours are sufficient while top-100 nearest neighbours provide a reliable upper bound covering all words.

| Words      | <i>Positive → Negative</i> |       |       |       |       |
|------------|----------------------------|-------|-------|-------|-------|
|            | 1750s                      | 1800s | 1850s | 1900s | 1950s |
| asylum     | 0.27                       | -0.24 | -0.54 | -0.52 | -0.65 |
| coronary   | 0.17                       | -0.13 | -0.22 | -0.15 | -0.68 |
| depressive | 0.3                        | -0.96 | -0.56 | -0.65 | -0.54 |
| germs      | 0.15                       | 0.26  | -0.14 | -0.55 | -0.61 |
| heartbreak | 0.18                       | -0.6  | -0.69 | -0.74 | -0.78 |

| Words       | <i>Negative → Positive</i> |       |       |       |       |
|-------------|----------------------------|-------|-------|-------|-------|
|             | 1750s                      | 1800s | 1850s | 1900s | 1950s |
| bloomers    | 0.01                       | -0.05 | 0.27  | 0.18  | 0.38  |
| destiny     | -0.54                      | 0.06  | 0.32  | 0.11  | 0.55  |
| dunk        | 0.42                       | 0.15  | None  | -0.18 | 0.19  |
| febrile     | -0.58                      | -0.53 | -0.66 | -0.54 | 0.33  |
| infatuation | -0.66                      | -0.63 | -0.52 | -0.35 | 0.40  |

Table 8: We present additional examples illustrating how the valence scores of words have shifted across centuries. The left column shows words that carried positive connotations historically but have since acquired negative ones, while the right column shows the inverse: words that were once negatively connoted but have taken on positive connotations over time.

| Words         | <i>Predominantly Negative</i> |       |       |       |       |
|---------------|-------------------------------|-------|-------|-------|-------|
|               | 1750s                         | 1800s | 1850s | 1900s | 1950s |
| afraid        | -0.35                         | -0.34 | -0.09 | -0.27 | -0.39 |
| angered       | -0.20                         | -0.58 | -0.66 | -0.66 | -0.69 |
| annihilation  | -0.49                         | -0.68 | -0.62 | -0.73 | -0.61 |
| bankruptcy    | -0.33                         | -0.37 | -0.46 | -0.55 | -0.46 |
| betray        | -0.52                         | -0.44 | -0.51 | -0.49 | -0.66 |
| chaos         | -0.27                         | -0.33 | -0.23 | -0.46 | -0.57 |
| stabbed       | -0.48                         | -0.68 | -0.74 | -0.60 | -0.56 |
| strangulation | -0.60                         | -0.47 | -0.67 | -0.69 | -0.26 |
| suicidal      | -0.27                         | -0.75 | -0.70 | -0.71 | -0.59 |

| Words      | <i>Predominantly Positive</i> |       |       |       |       |
|------------|-------------------------------|-------|-------|-------|-------|
|            | 1750s                         | 1800s | 1850s | 1900s | 1950s |
| abundant   | 0.28                          | 0.37  | 0.40  | 0.40  | 0.41  |
| enjoy      | 0.52                          | 0.60  | 0.62  | 0.63  | 0.86  |
| hugs       | -0.28                         | 0.34  | 0.56  | 0.56  | 0.40  |
| laughter   | 0.08                          | 0.33  | 0.49  | 0.49  | 0.75  |
| liking     | 0.35                          | 0.07  | 0.12  | 0.28  | 0.21  |
| lucky      | -0.14                         | 0.05  | 0.33  | 0.30  | 0.29  |
| marvel     | 0.03                          | 0.56  | 0.63  | 0.77  | 0.84  |
| merry      | 0.56                          | 0.71  | 0.76  | 0.79  | 0.54  |
| respectful | 0.45                          | 0.42  | 0.41  | 0.40  | 0.62  |

Table 9: Our VAD lexicons also capture words whose connotations have remained stable, largely positive or negative, over the 250 years of literary texts in Chronoberg, without undergoing any dramatic semantic shift.

examples include *awful* ( $-0.20 \rightarrow -0.66$ ), ( $-0.06 \rightarrow -0.28$ ). All 4 **H**-tier amelioration *Bastard* ( $+0.11 \rightarrow -0.48$ ), and *propaganda* words are also correctly detected (*awesome*:

| Word     | Positive → Negative |       |       |       |       |       |       |       |       |       |       |       |       |
|----------|---------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|          | 1750s               | 1770s | 1790s | 1810s | 1830s | 1850s | 1870s | 1890s | 1910s | 1930s | 1950s | 1970s | 1990s |
| asylum   | 0.16                | 0.06  | 0.16  | -0.01 | -0.20 | -0.50 | -0.49 | -0.54 | -0.52 | -0.54 | -0.46 | -0.10 | -0.10 |
| germs    | 0.30                | 0.28  | 0.21  | 0.28  | 0.31  | 0.17  | 0.15  | -0.54 | -0.52 | -0.68 | -0.53 | -0.53 | -0.06 |
| homeless | 0.20                | -0.23 | -0.44 | -0.53 | -0.66 | -0.65 | -0.64 | -0.66 | -0.63 | -0.49 | -0.52 | -0.52 | -0.20 |
| punk     | 0.17                | -0.31 | -0.16 | -0.16 | -0.16 | -0.21 | -0.21 | -0.16 | -0.12 | -0.35 | -0.14 | -0.11 | -0.11 |
| weird    | 0.20                | 0.20  | 0.05  | -0.11 | 0.00  | -0.32 | -0.29 | -0.26 | -0.36 | -0.47 | -0.44 | -0.17 | -0.17 |

Table 10: Valence trajectories for words undergoing a positive-to-negative shift, derived from lexicons built on Word2Vec models trained on 20-year temporal slices. Words that shift in the 50-year models show comparable transitions here, but the 20-year splits reveal a finer-grained progression of the valence scores than the coarser counterpart.

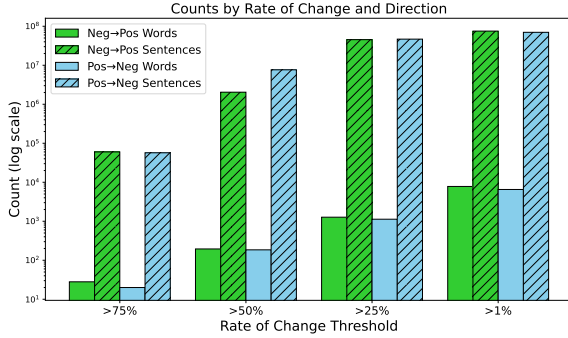


Figure 7: Distribution of valence shifts between consecutive 50-year intervals in Chronoberg. Bars show the count of words and sentences changing from positive to negative and vice versa, across different thresholds of rate of change. While most words remain effectively stable, thousands of samples exhibit substantial shifts in both directions.

$-0.78 \rightarrow -0.12$ ; *nice*:  $+0.18 \rightarrow +0.47$ ; *sophisticated*:  $+0.03 \rightarrow +0.27$ ; *Witch*:  $-0.30 \rightarrow -0.03$ . Medium-tier detections capture a further 11 words where the VAD signal is present but weaker, including the ameliorations *Minx* ( $-0.31 \rightarrow -0.12$ ), *terrific* ( $-0.61 \rightarrow -0.29$ ), *Vixen* ( $-0.20 \rightarrow -0.06$ ), and *maverick* ( $-0.12 \rightarrow +0.02$ ).

The 16 misrepresentations fall into two interpretable categories: (a) *Direction mismatch* ( $\ddagger$ ): Eleven words show corpus-level movement in the wrong direction. Words such as *bully* ( $-0.54 \rightarrow -0.18$ ) and *egregious* ( $-0.64 \rightarrow -0.36$ ) appear to ameliorate in book text because their dominant literary senses pull aggregate neighbourhood valence upward, masking the pejorative shift that occurred primarily in spoken and informal registers. Similarly, *nervous* ( $-0.52 \rightarrow -0.35$ ), *demagogue* ( $-0.52 \rightarrow -0.40$ ), and *Pervert* ( $-0.44 \rightarrow -0.34$ ) show small net positive shifts driven by oscillation around their already-negative baseline. *Batty*

( $-0.27 \rightarrow -0.01$ ) rises toward neutral in the literary corpus as the standard British English sense dominates, while the slur sense remained confined to informal registers. For Amelioration, *guy* ( $+0.17 \rightarrow -0.09$ ) is the sole miss: despite its modern colloquial positivity, the word carries slightly negative literary connotations across the full 250-year window, reflecting its historical association with grotesquely dressed effigies. Notably, *gay/gays* ( $+0.12 \rightarrow +0.52$ ) shows a strong positive trajectory (consistent with reclamation rather than pejoration).

(b) *Register gap* ( $\star$ ): Five words are assigned **L**ow because their pejorated senses were too infrequent in edited book text to produce measurable valence movement. *Fatal* ( $-0.62 \rightarrow -0.67$ ) was already strongly negative before 1750, leaving no trajectory to detect. *Weasel* ( $-0.05 \rightarrow +0.00$ ) and *Pig* ( $+0.02 \rightarrow +0.05$ ) remain near-neutral because the animal senses dominate published text. *Queen* ( $+0.42 \rightarrow +0.35$ ) and *Fairy* ( $+0.39 \rightarrow +0.46$ ) stay consistently positive throughout, as the regal and mythological senses respectively outweigh the pejorated senses in Chronoberg. *Sanction* ( $+0.18 \rightarrow +0.19$ ) is N/A: its two opposing meanings coexist and cancel, making VAD direction assessment inapplicable.

**Non-evaluative Changes.** Out of the 43 non-evaluative words, 8 correctly show flat valence, validating that the VAD lexicons do not generate spurious evaluative signals for non-evaluative changes. The words showing valence movement offer analytically interesting insights into how Non-evaluative change produces connotative drift.

*Metaphor*: Words such as *film* ( $-0.20 \rightarrow +0.14$ ) and *network* ( $+0.00 \rightarrow +0.17$ ) drift positive as their technological senses become dominant in later epochs. *Record* ( $+0.30 \rightarrow +0.11$ ), *stream* ( $+0.21 \rightarrow +0.09$ ), and *prop* ( $+0.19 \rightarrow -0.09$ )

| Word                           | Type       | $v_{1750}$ | $v_{1800}$ | $v_{1850}$ | $v_{1900}$ | $v_{1950}$ | VAD trend    | Tier | Agr. |
|--------------------------------|------------|------------|------------|------------|------------|------------|--------------|------|------|
| <i>Evaluative — Pejoration</i> |            |            |            |            |            |            |              |      |      |
| awful                          | Pejoration | -0.20      | -0.48      | -0.55      | -0.68      | -0.66      | Pejoration   | H    | ✓    |
| bastard                        | Pejoration | +0.11      | +0.25      | -0.05      | -0.35      | -0.48      | Pejoration   | H    | ✓    |
| fag                            | Pejoration | -0.15      | +0.09      | -0.12      | -0.16      | -0.51      | Pejoration   | H    | ✓    |
| hermaphrodite                  | Pejoration | +0.15      | -0.06      | +0.06      | +0.00      | -0.44      | Pejoration   | H    | ✓    |
| moron                          | Pejoration | +0.06      | —          | +0.22      | -0.59      | -0.26      | Pejoration   | H    | ✓    |
| propaganda                     | Pejoration | -0.06      | +0.00      | +0.14      | -0.24      | -0.28      | Pejoration   | H    | ✓    |
| slut                           | Pejoration | -0.08      | +0.00      | -0.33      | -0.56      | -0.50      | Pejoration   | H    | ✓    |
| twit                           | Pejoration | +0.35      | -0.46      | -0.51      | -0.55      | -0.25      | Pejoration   | H    | ✓    |
| batty                          | Pejoration | -0.27      | -0.10      | +0.04      | +0.14      | -0.01      | Amelioration | H    | ×    |
| bitch                          | Pejoration | +0.11      | +0.11      | +0.16      | +0.19      | -0.21      | Pejoration   | M    | ✓    |
| dyke                           | Pejoration | -0.09      | -0.14      | -0.14      | -0.28      | -0.33      | Pejoration   | M    | ✓    |
| prick                          | Pejoration | -0.31      | -0.18      | -0.30      | -0.45      | -0.41      | Pejoration   | M    | ✓    |
| queer                          | Pejoration | -0.24      | +0.00      | +0.16      | -0.14      | -0.27      | Pejoration   | M    | ✓    |
| bender                         | Pejoration | +0.05      | +0.06      | +0.00      | -0.20      | -0.01      | Pejoration   | M    | ✓    |
| faggot                         | Pejoration | -0.06      | -0.08      | +0.00      | -0.02      | -0.15      | Pejoration   | M    | ✓    |
| nigger                         | Pejoration | +0.08      | -0.05      | -0.32      | -0.36      | +0.03      | Pejoration   | M    | ✓    |
| bully                          | Pejoration | -0.54      | -0.62      | -0.55      | -0.51      | -0.18      | Amelioration | M    | ×    |
| demagogue                      | Pejoration | -0.52      | -0.11      | -0.44      | -0.16      | -0.40      | Amelioration | M    | ×    |
| dick                           | Pejoration | -0.10      | +0.05      | +0.16      | +0.10      | -0.01      | Amelioration | M    | ×    |
| egregious                      | Pejoration | -0.64      | -0.63      | -0.67      | -0.62      | -0.36      | Amelioration | M    | ×    |
| gay/gays                       | Pejoration | +0.12      | +0.07      | +0.10      | +0.52      | —          | Amelioration | M    | ×    |
| homo                           | Pejoration | +0.00      | —          | -0.14      | -0.01      | +0.04      | Amelioration | M    | ×    |
| nervous                        | Pejoration | -0.52      | -0.34      | -0.45      | -0.41      | -0.35      | Amelioration | M    | ×    |
| pervert                        | Pejoration | -0.44      | -0.62      | -0.79      | -0.79      | -0.34      | Amelioration | M    | ×    |
| pig                            | Pejoration | +0.02      | -0.13      | -0.13      | -0.09      | +0.05      | Amelioration | M    | ×    |
| fatal                          | Pejoration | -0.62      | -0.72      | -0.66      | -0.65      | -0.67      | Stable       | L    | ×    |
| weasel                         | Pejoration | -0.05      | -0.06      | -0.08      | -0.08      | +0.00      | Stable       | L    | ×    |
| queen                          | Pejoration | +0.42      | +0.45      | +0.42      | +0.41      | +0.35      | Stable       | L    | ×    |
| bent                           | Pejoration | -0.06      | -0.01      | -0.07      | -0.12      | -0.04      | Stable       | L    | ×    |
| fairy                          | Pejoration | +0.39      | +0.46      | +0.45      | +0.47      | +0.46      | Stable       | L    | ×    |

Table 11: Full validation results for in-scope evaluative words (Part 1a): Pejoration. Valence scores are per-epoch estimates from Chronoberg’s temporal VAD lexicons. Confidence tiers: **H**igh = strong multi-step signal ( $\geq 2$  steps,  $|\Delta_{\text{net}}| > 0.2$ ); **M**edium = signal present but weak or single-step; **L**ow = shift invisible to VAD lexicons (register gap or sub-threshold drift). ✓ = VAD direction agrees with expected annotated change; × = does not agree. ‘—’ = no data for that epoch.

decline in positivity as their associations with recordings, digital media, and theatrical props displace earlier neutral or positive uses. *Virus* ( $-0.34 \rightarrow -0.57$ ) shows a consistent negative drift as the pathological sense consolidates across the corpus window.

*Narrowing: Skyline* ( $+0.63 \rightarrow +0.02$ ) shows the sharpest single-epoch drop ( $-0.59$ ) as its meaning narrowed from a general horizon term to an urban architectural concept with more neutral connotations.

*Subjectification: Actually* ( $-0.01 \rightarrow +0.27$ ) and *starting* ( $-0.13 \rightarrow +0.16$ ) drift positive as they grammaticalise into affiliative discourse markers.

*Hopefully* ( $+0.02 \rightarrow -0.14$ ) drifts negative as its subjective “I hope” sense became associated with uncertainty and tentativeness.

*Multiple coexisting senses: Check* ( $-0.35 \rightarrow +0.21$ ) shows a dramatic positive shift as the confirmatory sense grew relative to the restraint/verification sense. *Stroke* ( $-0.40 \rightarrow -0.05$ ) ameliorates as positive senses (caress, brushstroke, stroke of luck) are represented alongside the medical sense.

*SemEval-2020 Stable control words* (Schlechtweg et al., 2020): Of the 24 control words, 9 are correctly classified as Stable (*bag, donkey, fiction, plane, player, risk, tip, tree, word*).

| Word                             | Type         | $v_{1750}$ | $v_{1800}$ | $v_{1850}$ | $v_{1900}$ | $v_{1950}$ | VAD trend    | Tier | Agr. |
|----------------------------------|--------------|------------|------------|------------|------------|------------|--------------|------|------|
| <i>Evaluative — Amelioration</i> |              |            |            |            |            |            |              |      |      |
| awesome                          | Amelioration | —          | −0.78      | −0.40      | −0.31      | −0.12      | Amelioration | H    | ✓    |
| nice                             | Amelioration | +0.18      | +0.54      | +0.68      | +0.57      | +0.47      | Amelioration | H    | ✓    |
| sophisticated                    | Amelioration | +0.03      | −0.29      | −0.34      | +0.23      | +0.27      | Amelioration | H    | ✓    |
| witch                            | Amelioration | −0.30      | −0.15      | −0.30      | −0.16      | −0.03      | Amelioration | H    | ✓    |
| minx                             | Amelioration | −0.31      | −0.25      | −0.18      | −0.28      | −0.12      | Amelioration | M    | ✓    |
| terrific                         | Amelioration | −0.61      | −0.55      | −0.51      | −0.49      | −0.29      | Amelioration | M    | ✓    |
| vixen                            | Amelioration | −0.20      | +0.16      | −0.18      | −0.06      | −0.06      | Amelioration | M    | ✓    |
| maverick                         | Amelioration | −0.12      | —          | +0.67      | +0.07      | +0.02      | Amelioration | M    | ✓    |
| guy                              | Amelioration | +0.17      | +0.04      | +0.02      | +0.15      | −0.09      | Pejoration   | H    | ×    |
| <i>Evaluative — Contronym</i>    |              |            |            |            |            |            |              |      |      |
| sanction                         | Contronym    | +0.18      | +0.22      | +0.23      | +0.30      | +0.19      | Stable       | M    | ×    |

Table 12: Full validation results for evaluative words (Part 1b): Amelioration and Contronym. See Table 11 for column definitions.

The remaining 15 show apparent movement, which is consistent with Chronoberg’s 250-year span: SemEval-2020 defines stability over a short two-timepoint window, so genuine long-horizon drift may account for some cases. The two strongest false positives are *quilt* (−0.15→+0.34) and *multitude* (−0.25→+0.04), both of which show upward drift likely driven by changes in genre distribution across epochs rather than genuine semantic change.

## B Analysis of Hate Speech and Harmful Language

This appendix section provides further information on how suitable hate-detection tools were identified and benchmarked, how they were applied on Chronoberg, and contains additional examples.

### B.1 Identifying Suitable Hate-detection Tools with HateCheck

We provide further insights into the choices underlying the main body’s hate-speech detection pipeline to contextualize harmful language in Chronoberg. In total, we have considered nine different modern hate-speech detection tools: Py-sentimiento toolkit (Pérez et al., 2021), Google Perspective API (Lees et al., 2022), Facebook RoBERTa (Vidgen et al., 2021), as well as 7 most popular Hugging Face models when filtering for the keyword “hate”. It is worth noting that all nine tools are built upon one of the two popular transformer architectures BERT (Devlin et al., 2019) and RoBERTa (Vidgen et al., 2021).

Consequently, we evaluated these tools on the HateCheck benchmark (Röttger et al., 2021), a suite of functional tests for hate-speech detection. As shown in Table 17, many existing hate-speech detection tools performed no better than random guessing. RoBERTa (Vidgen et al., 2021) was the only tool that stood out, demonstrating consistently strong performance, achieving high recall, while the remaining tools exhibited substantial limitations in either recall, precision, or overall reliability. Notably, Perspective API (Lees et al., 2022) achieved exceptionally high precision, making it particularly viable for curating a subset of potentially hateful sentences, despite its limited coverage.

To combine their strengths, we adopted a two-stage pipeline in the main body based on these HateCheck observations. First, we use RoBERTa (Vidgen et al., 2021) to flag a broad set of potentially hateful sentences. These sentences are then filtered by the Perspective API (Lees et al., 2022) to reduce false positives. This approach balances scalability and precision, addressing RoBERTa’s over-sensitivity and the Perspective API’s limited coverage.

### B.2 Sentences Containing Potential Hate in Chronoberg

We present a detailed analysis of Perspective API scores for sentences flagged as hateful by the RoBERTa model for Chronoberg. In Table 16(a), the distribution of these sentences across score intervals is shown, along with manual annotations of

| Word                              | Type      | $v_{1750}$ | $v_{1800}$ | $v_{1850}$ | $v_{1900}$ | $v_{1950}$ | VAD trend |
|-----------------------------------|-----------|------------|------------|------------|------------|------------|-----------|
| <i>Non-evaluative — Metaphor</i>  |           |            |            |            |            |            |           |
| attack                            | Metaphor  | -0.29      | -0.24      | -0.33      | -0.29      | -0.33      | Stable    |
| bug                               | Metaphor  | -0.31      | -0.29      | -0.35      | -0.33      | -0.24      | Stable    |
| edge                              | Metaphor  | -0.08      | -0.04      | -0.05      | -0.07      | -0.13      | Stable    |
| land                              | Metaphor  | +0.18      | +0.13      | +0.18      | +0.16      | +0.14      | Stable    |
| platform                          | Metaphor  | +0.03      | +0.06      | +0.12      | +0.10      | +0.08      | Stable    |
| bit                               | Metaphor  | +0.14      | -0.13      | -0.15      | -0.13      | -0.19      | Shifting  |
| broadcast                         | Metaphor  | +0.09      | +0.13      | +0.06      | -0.05      | +0.07      | Shifting  |
| cell                              | Metaphor  | -0.06      | +0.01      | -0.01      | -0.16      | -0.04      | Shifting  |
| cloud                             | Metaphor  | +0.04      | -0.08      | -0.05      | -0.14      | +0.06      | Shifting  |
| film                              | Metaphor  | -0.20      | -0.04      | +0.06      | +0.14      | +0.14      | Shifting  |
| mouse                             | Metaphor  | -0.17      | -0.07      | -0.05      | -0.03      | +0.02      | Shifting  |
| network                           | Metaphor  | +0.00      | -0.03      | +0.08      | +0.01      | +0.17      | Shifting  |
| prop                              | Metaphor  | +0.19      | +0.18      | +0.14      | -0.01      | -0.09      | Shifting  |
| rag                               | Metaphor  | +0.06      | -0.07      | -0.05      | -0.09      | -0.12      | Shifting  |
| record                            | Metaphor  | +0.30      | +0.32      | +0.20      | +0.15      | +0.11      | Shifting  |
| stab                              | Metaphor  | -0.48      | -0.49      | -0.50      | -0.39      | -0.49      | Shifting  |
| stream                            | Metaphor  | +0.21      | +0.08      | +0.13      | +0.14      | +0.09      | Shifting  |
| tablet                            | Metaphor  | +0.12      | +0.05      | -0.12      | -0.08      | +0.01      | Shifting  |
| thump                             | Metaphor  | -0.19      | -0.38      | -0.44      | -0.42      | -0.28      | Shifting  |
| virus                             | Metaphor  | -0.34      | -0.52      | -0.44      | -0.38      | -0.57      | Shifting  |
| web                               | Metaphor  | -0.01      | -0.01      | +0.03      | +0.02      | -0.10      | Shifting  |
| <i>Non-evaluative — Narrowing</i> |           |            |            |            |            |            |           |
| car                               | Narrowing | +0.21      | +0.16      | +0.16      | +0.13      | +0.10      | Shifting  |
| diet                              | Narrowing | +0.13      | +0.00      | +0.19      | +0.22      | +0.13      | Shifting  |
| estate                            | Narrowing | +0.34      | +0.35      | +0.33      | +0.19      | +0.31      | Shifting  |
| skyline                           | Narrowing | —          | +0.63      | +0.04      | +0.05      | +0.02      | Shifting  |

Table 13: Full validation results for non-evaluative and control words (Part 2a). The “Type” column indicates whether the change in the meaning of the words falls into one of these categories: “Metaphor”, “Narrowing”, “Generalization”, “Subjectification”, or “Multiple coexisting senses” (“Polysemy” or “Domain shift”). “VAD Trend” being “Stable” indicates valence correctly stays flat; “Shifting” indicates an unexpected directional shift. Within each subcategory, stable words precede shifting words. (*Continued in Parts 2b–2c.*)

100 sampled sentences per range to estimate the precision. No manual revision was conducted for scores below 0.5, as the [0.5, 0.6) range already yielded 16 false positives, indicating a substantial drop in precision. In consequence, lower thresholds seem to be impractical for reliable hate-speech filtering. Table 16(b) complements this by showing how varying the Perspective API threshold affects both the number of flagged sentences and the estimated precision. We extrapolated precision estimates from the interval-level annotations, due to the infeasibility of reviewing all 3.3 million samples. Notably, 2.4 million sentences, making up 72.1%, that were labelled as hateful by RoBERTa scored below 0.1 in Perspective API, indicating a high false positive rate. Despite this, RoBERTa exhibited strong recall during evaluation and likely also captured a majority of hateful sentences, albeit imprecisely. Using a threshold of 0.7 with Perspective API resulted in a highly precise subset of 5,597 sentences.

We observed that while varying the Perspective

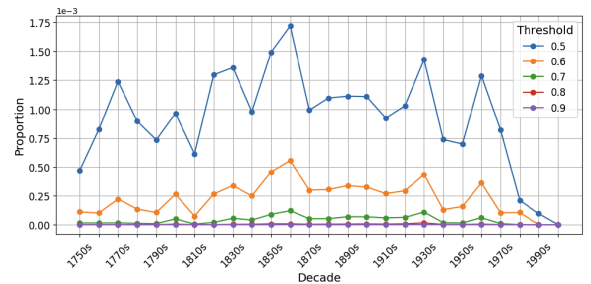


Figure 8: Proportion of hateful sentences over time for different Perspective API thresholds

API threshold influences the volume of hate detected across decades, as seen in Figure 8, the overall temporal patterns remain similar and are retained. In comparison with the distribution as flagged by RoBERTa, more substantial differences are revealed, particularly in earlier historical periods, whereas Perspective API consistently detects less hate across all thresholds.

| Word                                               | Type              | $v_{1750}$ | $v_{1800}$ | $v_{1850}$ | $v_{1900}$ | $v_{1950}$ | VAD trend |
|----------------------------------------------------|-------------------|------------|------------|------------|------------|------------|-----------|
| <i>Non-evaluative — Generalization</i>             |                   |            |            |            |            |            |           |
| head                                               | Generalization    | −0.02      | −0.04      | +0.02      | −0.02      | −0.05      | Stable    |
| circle                                             | Generalization    | +0.01      | +0.13      | +0.08      | +0.07      | +0.08      | Shifting  |
| lass                                               | Generalization    | +0.19      | +0.33      | +0.22      | +0.11      | −0.10      | Shifting  |
| <i>Non-evaluative — Subjectification</i>           |                   |            |            |            |            |            |           |
| major                                              | Subjectification  | +0.09      | +0.10      | +0.09      | +0.10      | +0.06      | Stable    |
| actually                                           | Subjectification  | −0.01      | +0.09      | +0.21      | +0.10      | +0.27      | Shifting  |
| headed                                             | Subjectification  | −0.15      | +0.03      | −0.04      | +0.04      | −0.03      | Shifting  |
| hopefully                                          | Subjectification  | +0.02      | +0.09      | +0.25      | −0.01      | −0.14      | Shifting  |
| promise                                            | Subjectification  | +0.25      | +0.34      | +0.27      | +0.26      | +0.37      | Shifting  |
| started                                            | Subjectification  | −0.19      | +0.01      | +0.02      | −0.04      | +0.01      | Shifting  |
| starting                                           | Subjectification  | −0.13      | −0.04      | −0.05      | +0.09      | +0.16      | Shifting  |
| <i>Non-evaluative — Multiple coexisting senses</i> |                   |            |            |            |            |            |           |
| computer                                           | Domain/Technology | —          | —          | +0.12      | +0.16      | +0.19      | Stable    |
| call                                               | Polysemy          | −0.01      | +0.12      | +0.22      | +0.21      | +0.19      | Shifting  |
| calls                                              | Polysemy          | +0.04      | −0.03      | −0.03      | −0.05      | +0.12      | Shifting  |
| check                                              | Polysemy          | −0.35      | −0.31      | −0.25      | +0.21      | +0.21      | Shifting  |
| democrat                                           | Domain/Technology | +0.02      | −0.08      | +0.03      | +0.04      | +0.07      | Shifting  |
| monitor                                            | Polysemy          | +0.02      | +0.22      | −0.05      | −0.11      | +0.13      | Shifting  |
| privacy                                            | Domain/Technology | +0.21      | +0.09      | +0.15      | +0.14      | +0.25      | Shifting  |
| stroke                                             | Polysemy          | −0.40      | −0.34      | −0.24      | −0.11      | −0.05      | Shifting  |

Table 14: Full validation results for non-evaluative words, continued (Part 2b). See Table 13 for column definitions. (Continued in Part 2c.)

| Word                       | Type   | $v_{1750}$ | $v_{1800}$ | $v_{1850}$ | $v_{1900}$ | $v_{1950}$ | VAD trend |
|----------------------------|--------|------------|------------|------------|------------|------------|-----------|
| <i>SemEval-2020 Stable</i> |        |            |            |            |            |            |           |
| bag                        | Stable | +0.01      | +0.04      | +0.04      | +0.05      | +0.05      | Stable    |
| donkey                     | Stable | +0.01      | −0.01      | +0.08      | +0.03      | +0.07      | Stable    |
| fiction                    | Stable | +0.36      | +0.37      | +0.33      | +0.30      | +0.40      | Stable    |
| plane                      | Stable | −0.02      | −0.02      | +0.03      | +0.05      | −0.01      | Stable    |
| player                     | Stable | +0.30      | +0.26      | +0.31      | +0.30      | +0.29      | Stable    |
| risk                       | Stable | −0.26      | −0.33      | −0.27      | −0.37      | −0.35      | Stable    |
| tip                        | Stable | −0.08      | −0.14      | −0.10      | −0.07      | −0.14      | Stable    |
| tree                       | Stable | +0.11      | +0.06      | +0.11      | +0.10      | +0.06      | Stable    |
| word                       | Stable | +0.15      | +0.14      | +0.05      | +0.09      | +0.16      | Stable    |
| ball                       | Stable | −0.20      | +0.03      | +0.09      | +0.03      | +0.08      | Shifting  |
| chairman                   | Stable | +0.11      | +0.24      | +0.20      | +0.21      | +0.20      | Shifting  |
| contemplation              | Stable | +0.10      | +0.37      | +0.35      | +0.34      | +0.37      | Shifting  |
| face                       | Stable | −0.14      | +0.00      | +0.05      | +0.00      | +0.00      | Shifting  |
| gas                        | Stable | −0.12      | +0.01      | −0.03      | −0.06      | −0.17      | Shifting  |
| graft                      | Stable | +0.10      | +0.15      | +0.12      | −0.64      | +0.15      | Shifting  |
| lane                       | Stable | +0.06      | +0.17      | +0.20      | +0.20      | +0.11      | Shifting  |
| multitude                  | Stable | −0.25      | −0.13      | −0.06      | −0.10      | +0.04      | Shifting  |
| ounce                      | Stable | +0.15      | +0.06      | +0.06      | +0.05      | −0.02      | Shifting  |
| part                       | Stable | +0.04      | +0.06      | +0.15      | +0.13      | +0.15      | Shifting  |
| pin                        | Stable | −0.11      | −0.13      | −0.02      | −0.04      | +0.03      | Shifting  |
| quilt                      | Stable | −0.15      | +0.18      | +0.16      | +0.15      | +0.34      | Shifting  |
| relationship               | Stable | +0.47      | +0.37      | +0.42      | +0.37      | +0.28      | Shifting  |
| savage                     | Stable | −0.48      | −0.54      | −0.56      | −0.56      | −0.63      | Shifting  |
| twist                      | Stable | −0.14      | −0.13      | −0.11      | −0.20      | −0.25      | Shifting  |

Table 15: Full validation results for control words (Part 2c). Words from SemEval-2020 (Schlechtweg et al., 2020) annotated as “Stable” serve as a control group. Of the 24 words, 9 are correctly classified as Stable and 15 show apparent valence movement; some of the latter may reflect genuine long-horizon drift below SemEval’s short two-timepoint detection threshold rather than classifier noise. See Table 13 for column definitions.

| Score $\in$ | Sentences | TP  | FP | Score $\geq$ | Sentences | P     |
|-------------|-----------|-----|----|--------------|-----------|-------|
| [0.0, 0.1)  | 2,411,275 | -   | -  | 0.0          | 3,343,433 | -     |
| [0.1, 0.2)  | 360,085   | -   | -  | 0.1          | 932,158   | -     |
| [0.2, 0.3)  | 228,128   | -   | -  | 0.2          | 572,073   | -     |
| [0.3, 0.4)  | 148,038   | -   | -  | 0.3          | 343,945   | -     |
| [0.4, 0.5)  | 99,557    | -   | -  | 0.4          | 195,907   | -     |
| [0.5, 0.6)  | 68,540    | 84  | 16 | 0.5          | 96,350    | 87.4% |
| [0.6, 0.7)  | 22,213    | 95  | 5  | 0.6          | 27,810    | 95.8% |
| [0.7, 0.8)  | 5,116     | 99  | 1  | 0.7          | 5,597     | 99.1% |
| [0.8, 0.9)  | 470       | 100 | 0  | 0.8          | 481       | 100%  |
| [0.9, 1.0]  | 11        | 11  | 0  | 0.9          | 11        | 100%  |

(a) Distribution and precision across score intervals

(b) Precision and coverage at varying thresholds

Table 16: Perspective API score analysis for sentences flagged as hateful by RoBERTa.

| Model                                             | Acc          | F1           | P            | R            |
|---------------------------------------------------|--------------|--------------|--------------|--------------|
| Perspective API (Lees et al., 2022)               | 0.578        | 0.559        | <b>0.993</b> | 0.389        |
| pysentimiento (Pérez et al., 2021)                | 0.521        | 0.527        | 0.820        | 0.388        |
| Facebook’s RoBERTa (Vidgen et al., 2021)          | <b>0.956</b> | <b>0.968</b> | 0.963        | <b>0.973</b> |
| English Abusive MuRIL (Das et al., 2022)          | 0.491        | 0.558        | 0.694        | 0.466        |
| BERT HateXplain (Mathew et al., 2021)             | 0.384        | 0.270        | 0.730        | 0.165        |
| DehateBERT Mono English (Aluru et al., 2020)      | 0.425        | 0.351        | 0.784        | 0.226        |
| IMSyPP Hate Speech (Kralj Novak et al., 2022)     | 0.750        | 0.826        | 0.790        | 0.866        |
| Twitter RoBERTa Large Hate (Antypas et al., 2023) | 0.615        | 0.640        | 0.898        | 0.497        |
| DistilRoBERTa Hateful Speech (badmatr11x, 2023)   | 0.568        | 0.652        | 0.730        | 0.590        |

Table 17: Performances of models on full HateCheck test set. The best values are indicated in bold.

### B.3 Further qualitative examples of harmful language analysis

In complement to the examples shown in section 3.2 of the main body, we present additional examples examining the alignment between hate-check outputs from LLMs and VAD scores in Table 18. We note here that for the hate-speech experiments, we queried the Perspective API and OpenAI Moderation API via their respective public endpoints; no local GPU inference was conducted for these tools. Whereas, the four LLMs (Mistral-8x7B, Qwen2.5-72B, GPT-OSS-120B, DeepSeek-R1-70B) were hosted locally on 4 A6000 Nvidia GPUs. No fine-tuning of any model was performed, and each run classifying all the target sentences accounted for only  $\sim 1$  GPU hour. We observe comparable trends and discrepancies in sentiment classification, particularly in how modern classifiers often fail to recognize historically situated expressions of hate or, conversely, overgeneralize from present-day connotations.

| YEAR  | Sentences                                                                                                | Hate-Check Models |        |      |         | Valence Score | Affective Sentiment |
|-------|----------------------------------------------------------------------------------------------------------|-------------------|--------|------|---------|---------------|---------------------|
|       |                                                                                                          | RoBERTa+Persp     | OpenAI | Qwen | Mistral |               |                     |
| 1800s | We has slaves too; We has n*ggers to a stand-still.                                                      | 🚩                 | 🚩      | 🚩    | 🚩       | -0.37         | 👎                   |
| 1850s | I hate women.                                                                                            | 🚩                 | 🚩      | 🚩    | 🚩       | -0.40         | 👎                   |
| 1900s | You never want to take a n*gger into your conferences.                                                   | 🚩                 | 🚩      | 🚩    | 🚩       | -0.18         | 👎                   |
| 1900s | Kill every black bastard befo mornin!                                                                    | 🚩                 | 🚩      | 🚩    | 🚩       | -0.36         | 👎                   |
| 1900s | The Bhutanese women are the ugliest specimens of femininity I have ever seen.                            | 🚩                 | 🚩      | 🚩    | 🚩       | 0.45          | 👎                   |
| 1950s | I hate chicago, i hate americans!                                                                        | 🚩                 | 🚩      | 🚩    | 🚩       | -0.40         | 👎                   |
| 1950s | I hate the germans!                                                                                      | 🚩                 | 🚩      | 🚩    | 🚩       | -0.40         | 👎                   |
| 1750s | A man may play with decency; but if he games, he is disgraced.                                           | ✓                 | ✓      | ✓    | ✓       | -0.68         | 👎                   |
| 1750s | Defamation and calumny never attack, where there is no weak place; they magnify, but they do not create. | ✓                 | ✓      | ✓    | ✓       | -0.76         | 👎                   |
| 1750s | Thou traitor, hie away; By all my stars I thou enviest Tom Thumb                                         | ✓                 | 🚩      | 🚩    | 🚩       | -0.69         | 👎                   |
| 1800s | Who the beggar was that i killed                                                                         | ✓                 | ✓      | 🚩    | ✓       | -0.60         | 👎                   |
| 1800s | What hatred she distills!                                                                                | ✓                 | ✓      | 🚩    | 🚩       | -0.72         | 👎                   |
| 1850s | The piece was stupid beyond expression                                                                   | ✓                 | ✓      | ✓    | ✓       | -0.57         | 👎                   |
| 1950s | So it is a hell of women, is it?                                                                         | ✓                 | ✓      | 🚩    | ✓       | -0.35         | 👎                   |
| 1750s | The conversation at supper was very gay.                                                                 | 🚩                 | ✓      | ✓    | ✓       | 0.37          | 👍                   |
| 1750s | In my way home to my tent, I saw a faggot lying in the way,                                              | 🚩                 | 🚩      | 🚩    | 🚩       | 0.05          | 👍                   |
| 1850s | Religion prescribes obedience.                                                                           | 🚩                 | ✓      | ✓    | ✓       | 0.08          | 👍                   |
| 1900s | I may cut you out of my gold expedition, if you get gay.                                                 | 🚩                 | 🚩      | 🚩    | 🚩       | 0.17          | 👍                   |

Table 18: Extended illustrative examples. Comparison of sentence-level classifications in Chronoberg across time intervals using LLM-based hate-check tools (RoBERTa+Perspective API, OpenAI, Qwen, and Mistral) (🚩- Hate, ✓- Non-hate) and valence-based scoring (👎- Negative, 👍- Positive sentiment). We grouped the instances based on model agreement. The first group of rows illustrates cases where all tools collectively classify an instance as harmful. The group of rows 2-3 shows instances where they disagree, classifying them as either positive or negative. While VAD lexicons provide interpretable complementary signals, we recognize that harmful texts are inherently subjective; thus, we do not regard them as definitive solutions to LLM misclassification but as potential tools to enhance LLM performance. Note, we have partially masked offensive N-words with asterisks.

## C Chronoberg Datasheet

### Motivation

**For what purpose was the dataset created?** (e.g. was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)

The Chronoberg dataset was created to provide a temporally structured corpus to support diachronic linguistic analysis and the evaluation of LLMs under temporal shift. While existing resources provide broad coverage, they typically lack long-term temporal structure, and are not well-suited to studying semantic drift and diachronic variation. Chronoberg was designed to support tasks such as:

- Analysis of detection of discriminatory and sensitive language in historical contexts.
- Construction of historically grounded affective lexicons for systematic linguistic and affective analysis, and
- Evaluation of temporal generalization.

**What (other) tasks could the dataset be used for?** Are there obvious tasks for which it should not be used?

The dataset was created to provide a scalable benchmark for tasks such as:

- Concept drift modelling and continual learning,
- Editing and unlearning to modify or update interpretations of certain English words or sentences.

The VAD score annotations in the dataset are NOT intended for tasks that require absolute semantic evaluation, but study relative semantic change.

**Who created this dataset (e.g., which team, research group) and on behalf of which entity?** (e.g., company, institution, organization)

The creators of the dataset are: Niharika

Hegde<sup>1,4\*</sup>, Subarnaduti Paul<sup>2\*</sup>, Lars Joel-Frey<sup>4</sup>, Manuel Brack<sup>1,3†</sup>, Kristian Kersting<sup>1,3,4,5</sup>, Martin Mundt<sup>2†</sup>, Patrick Schramowski<sup>1,3,4,6†</sup>. The associated affiliations are: <sup>1</sup>German Research Center for Artificial Intelligence (DFKI), Darmstadt, Germany, <sup>2</sup>Department of Computer Science and

Mathematics, University of Bremen, Bremen, Germany, <sup>3</sup>Hessian Center for AI (hessian.AI), Darmstadt, Germany, <sup>4</sup>Department of Computer Science, TU Darmstadt, Darmstadt, Germany, <sup>5</sup>Centre for Cognitive Science, TU Darmstadt, Darmstadt, Germany and <sup>6</sup>CERTAIN, Germany.

**Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

No funding was received for the creation of the dataset.

**Any other comments?**

No further comments.

### Composition

**What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances in the Chronoberg dataset represent digitized books sourced from Project Gutenberg. Each instance corresponds to:

- Full text of a book
- Temporal metadata (publication year)
- Lexicons that capture affective sentiment dimensions such as Valence, Dominance, and Arousal.
- Sentence level annotations associating a VAD score.

**How many instances are there in total (of each type, if appropriate)?**

Chronoberg is composed of 2.7 billion tokens, representing roughly 91 million sentences from 25,061 English-language books published between 1750 and 2000, with additional metadata in the form of temporally-aligned VAD lexicons that span 335,804 words.

**Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please

describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Chronoberg is curated from Project Gutenberg, an openly accessible corpus of literary texts. Our primary focus was to derive a temporally structured dataset suitable for studying diachronic semantic drift, i.e., how word and sentence meanings evolve across centuries. To this end, we restricted our scope to Late Modern English works published between 1750 and 2000. Since Project Gutenberg’s metadata on original publication date is often inaccurate or missing, we developed an inference pipeline that leverages the OpenLibrary API to obtain corrected publication years. After this process, we retained a total of 25,061 books.

**What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features?** In either case, please provide a description.

Each instance represents processed texts. In addition, we have also introduced temporally-aligned VAD lexicons as part of our metadata, where each instance represents processed English words.

**Is there a label or target associated with each instance?** If so, please provide a description.

The dataset is grouped by year of publication. So, texts contain labels in the form of a specific year to which they belong. In addition, we have provided “valence”, “arousal”, and “dominance” annotations for each sentence in a specific time period.

**Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

There is no information missing, as we have excluded such examples from the dataset in its construction process.

**Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

Relationships between the individual instances are made explicit through their shared temporal alignment. Each book instance is linked to a publication year, which allows for grouping, comparison and sequential ordering across time. Instances can also be grouped into genres such as Fiction, History, Social Life, Conduct of Life, and Travel, inherited from the original Project Gutenberg corpus. At the same time, given the historical span of 1750–2000, the texts could also be categorized with respect to major historical contexts and events of the period. We did not pursue this latter categorization in depth, as we believe it requires domain expertise beyond the scope of our work.

**Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

Yes. Although the best performing year predictor yields an uncertainty of 3-5 years, this margin is negligible at the scale of our diachronic analysis. To ensure robustness against temporal noise and to better capture semantic shifts over time, we recommend that alternately created training, validation and test sets be constructed within coarse temporal bins no smaller than 15 years.

**Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

Yes, as mentioned above there is an uncertainty of around 3-5 years as a consequence of our publication date inference pipeline. The uncertainty originates from the fact that publication dates are often missing or refer to the time of digitization in their original online repository and alternate sources needed to be inquired. We believe the uncertainty is acceptable in light of our considered 250 year time range.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and

any restrictions associated with them, as well as links or other access points, as appropriate.

Yes, the dataset is self-contained. There are no access restrictions or required external resources.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?** If so, please provide a description.

No.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.

Yes. We acknowledge that the literary texts from the 1750s to the 2000s that compose Chronoberg may contain sensitive content that could be offensive, insulting, or threatening to certain groups. While we do not intend to objectify anyone, we also aim to preserve the integrity of the original works without alteration and avoiding historical erasure. At the same time, this approach opens up new possibilities for exploring how emotions were directed toward specific groups during different historical periods.

**Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

Yes, we acknowledge that Chronoberg contains instances of text that may refer to groups of people or individuals, either directly or indirectly. As it comprises historical literary works from the 1750s to the 2000s, some texts are curated from literal biographies of historical figures, while others relate to wars and other significant events of the period.

**Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Yes, there are literary works that refer to or identify subpopulations by age, group or gender, considering that literary works span across different genres and include historical content.

**Is it possible to identify individuals (i.e., one or more natural persons), either di-**

**rectly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.

Yes, in particular historical figures are explicitly identified in books containing historical non-fiction content.

**Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

Yes. While the dataset is composed of historical texts rather than personal records, it contains language that may be considered sensitive. This includes expressions of racial, ethnic, religious, gendered, political bias reflective of the time periods covered (1750-2000). Such content may involve discriminatory or offensive terminology, depiction of marginalized groups or outdated normative assumptions. However, the dataset itself does not contain personal identifiers, financial, health, biometric or government identification data.

**Any other comments?**

No further comments.

### Collection Process

**How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was directly acquired from Project Gutenberg, an online library of copyright-free e-books from the past few centuries. It also allows easy access by mirroring their entire catalogue or downloading their e-book collection via an API. The data was directly observable as raw texts.

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation,**

**software program, software API)? How were these mechanisms or procedures validated?**

We have followed the official recommendation from Project Gutenberg to download the RDF files of the books via mirroring. The official mirror links can be found at their official website <https://www.gutenberg.org/help/mirroring.html>. We have also used their official repository <https://github.com/gutenbergtools> to interact with their resources when needed.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

We have specifically focused on curating English texts from the period 1750–2000. However, in some cases, the original publication date of a work was either missing or inaccurately recorded. To address this, we employed an additional sampling strategy to ensure that works were properly curated and accurately categorized into their respective time intervals.

**Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

Only the authors and co-authors were responsible for the collection of the data.

**Over what time-frame was the data collected? Does this time-frame match the creation time-frame of the data associated with the instances (e.g., recent crawl of old news articles)?** If not, please describe the time-frame in which the data associated with the instances was created.

The data was curated between October 2024 and March 2025. Since the dataset consists solely of historically published literary works, its content remains unaffected by the timeline of collection and compilation.

**Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No

**Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

Given its historical context, the data can relate to people or groups of individuals. The literary texts constituting Chronoberg represent biographies, works on important historical events between 1750 and 2000, social life, and similar aspects pertaining to historical society.

**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

No, the data was not collected directly from any individual; rather, it was acquired from Project Gutenberg, a historical corpus of literary works, some of which may pertain to certain individuals.

**Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No.

**Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

The literary works obtained from Project Gutenberg are copyright-free and freely distributable. We have ensured that no information in these works was altered, preserving their integrity and originality. Given the historical time-frame covered by Chronoberg, the individuals referenced in these texts are not available to provide consent.

**If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable.

**Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the out-

comes, as well as a link or other access point to any supporting documentation.

No such analysis with respect to data protection or privacy could be conducted as a) historical figures have long been deceased, b) information on historical figures has been disseminated in various historical works throughout time, c) Project Gutenberg has already provided a curated public archive that has excluded copy-righted and non-consensual material outside the public domain. Chronoberg derives itself from Project Gutenberg and thus not induce any new impact regarding data subjects.

### Any other comments?

No further comments.

### Preprocessing/cleaning/labelling

**Was any preprocessing/cleaning/labelling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remainder of the questions in this section.

We have followed several steps of preprocessing and labelling of our curated raw texts from Project Gutenberg.

- **Determination of original publication date:** We observed that e-books in Project Gutenberg often lack accurate publication years, or in some cases, the information is entirely missing. This step is particularly important for Chronoberg, as incorrect publication years would lead to misclassification of texts into the wrong temporal intervals.
- **Data partitioning:** The texts are grouped by year, resulting in 250 separate splits corresponding to individual years. In addition, we created broader bins spanning 50-year intervals.
- **Data filtering:** We removed all non-alphanumeric characters from the texts to facilitate easier adaptation across various downstream applications.

**Was the “raw” data saved in addition to the preprocessed/cleaned/labelled data (e.g., to support unanticipated future uses)?** If

so, please provide a link or other access point to the “raw” data.

Yes, as a direct part of the dataset, we have made available two versions of Chronoberg: one consisting of the raw texts grouped by publication year, and another with processed texts, split into annotated sentences and likewise organized by publication year.

**Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.

Yes, a public GitHub repository explaining the pre-processing, cleaning, and labelling process is available at: <https://github.com/paulsubarna/Chronoberg/>. Simultaneously, the dataset is available on HuggingFace: <https://huggingface.co/datasets/spaul25/Chronoberg>

### Any other comments?

No further comments.

### Uses

**Has the dataset been used for any tasks already?** If so, please provide a description.

The dataset has not been publicly available before. We highlight several potential downstream applications of Chronoberg in our accompanying experimental analysis, including: (i) validating temporally calibrated VAD lexicons against words with documented evaluative and non-evaluative changes from literature, (ii) inspecting words and sentences within Chronoberg that have undergone diachronic shifts and outline future prospects, such as (iii) unlearning or modifying specific connotations in words used in contemporary English texts.

**Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

No.

**What (other) tasks could the dataset be used for?**

Beyond the experiments presented in this paper, Chronoberg supports culturomics research, temporally grounded sentiment analysis and furthers research of historically situated NLP benchmarks. The VAD lexicons may also support research on social bias and stereotyping across historical periods. As part of our metadata, we also provide valence, arousal, and dominance scores for each sentence

in Chronoberg, aiming to capture the affective sentiment expressed in the text. However, given the sensitivity of certain sentences, we strongly caution against interpreting these scores as definitive labels of positivity or negativity.

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labelled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

Alongside their publication year, we have also annotated the sentences in Chronoberg based on their VAD scores, denoting their affective polarity. However, we acknowledge that not every negatively scored sentence based on the VAD score can be used as a harmful sentence. So, we encourage users to not treat the dataset for benchmarking hateful vs non-hateful applications. Similarly, as this data may directly or indirectly involve individuals, we discourage its use to specifically single out any individuals solely based on the VAD lexicons and affective polarity scores of the texts.

**Are there tasks for which the dataset should not be used?** If so, please provide a description.

Please refer to the previous question.

**Any other comments?**

No further comments.

#### Distribution

**Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

Yes, the dataset is publicly available at HuggingFace <https://huggingface.co/datasets/spaul25/Chronoberg>, and the source code is available at GitHub <https://github.com/paulsubarna/Chronoberg/>

**How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?**

We have made the dataset publicly available at HuggingFace: <https://huggingface.co/datasets/spaul25/Chronoberg>

**When will the dataset be distributed?**

The dataset will be distributed after de-anonymization.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is distributed under the BSD 2-Clause “Simplified” License. It also comes under the full list of Project Gutenberg licenses, which can be found in <https://www.gutenberg.org/policy/license.html>

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

**Any other comments?**

No further comments.

#### Maintenance

**Who will be supporting/hosting/maintaining the dataset?**

The authors will be responsible for the maintenance and support of the dataset.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

The creators of the datasets: Subarnaduti Paul and Niharika Hegde can be contacted at [spaul@uni-bremen.de](mailto:spaul@uni-bremen.de) and [niharika.hegde@dfki.de](mailto:niharika.hegde@dfki.de) respectively.

**Is there an erratum?** If so, please provide a link or other access point.

There currently exists no erratum.

**Will the dataset be updated (e.g., to correct labelling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Currently, no immediate updates are envisioned. If there appears to be an urgent need to update the dataset, the authors will be responsible for uploading a new version. We will update the version at HuggingFace and communicate the news on our website and HuggingFace.

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

As the acquired data is copyright-free and free to distribute, we don't believe there is any applicable limits on the retention of the data associated with the instances.

**Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes, we will continue to host and support older versions of the dataset. HuggingFace, as a platform, supports versioning.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

The code for data generation is publicly available on GitHub at <https://github.com/paulsubarna/Chronoberg>. Validation/Verification of future contributions will not be in the scope of the authors. Yes, there are numerous possibilities to extend the dataset, such as:

- Extending beyond the 1750–2000 time-frame, the corpus can also be viewed as an ever-growing temporal dataset of historical contexts, given copy-right laws being based on 20 and 100 year cut-offs respectively.
- The dataset can be extended with several other languages besides Late Modern English.

### **Any other comments?**

No further comments.