EmoTracker—A New Framework for Modeling and Forecasting Diachronic Emotion Dynamics

Max Tiessler[1,2], Quim Motger[1], Florina Piroi[2], and Andreas Baumann[3,4]

[1] Departament d'Enginyeria de Serveis i Sistemes d'Informació, Universitat Politècnica de Catalunya, Barcelona, Spain
[2] Institute of Information Systems Engineering, TU Wien, Vienna, Austria
[3] Department of German Studies, University of Vienna, Vienna, Austria
[4] Research Network Data Science, University of Vienna, Vienna, Austria

Abstract

Existing computational approaches to diachronic semantics and emotion analysis typically study word meaning change and emotional evolution separately, limiting our understanding of how emotions evolve in proportion to the sense level. To bridge this gap, we propose EmoTracker, a novel framework that integrates diachronic sense modeling with Valence-Arousal-Dominance (VAD) emotion tracking to model and predict temporal emotion-sense trajectories. Our contribution is threefold. First, we develop a method for constructing temporal emotion datasets by integrating diachronic sense data with three different VAD lexicons. Second, we implement an LSTM architecture with attention mechanisms and momentum-based features to forecast emotional trajectories over time. Third, we provide interactive 3D visualizations to explore emotion dynamics over time, and 4D visualizations to capture the diachronic joint evolution of emotions and senses in the VAD space. Our evaluation shows that, among the selected lexicons, NRC-VAD is the most suitable for temporal modeling, though it also reveals the challenges in modeling dominance across lexicons. EmoTracker bridges diachronic semantics and emotion analysis, providing a comprehensive framework for computational humanities research.

Keywords: diachronic emotion analysis, LSTM forecasting, temporal sentiment analysis, VAD modeling, semantic change modeling

1 Introduction

Language is constantly evolving, shaped by cultural, social, and technological changes. This is particularly visible in the lexicon, as meanings and emotional connotations of words change over time. Such changes typically happen when words become more polysemous by obtaining now senses or when they lose some of their senses. For example, the English word *gay* once primarily meant 'cheerful' with a clearly positive connotation but today more neutrally also refers to sexual orientation; *awful* originally just meant 'awe-inspiring', but now mostly carries a clearly more negative connotation ('disgusting') [9; 15]. Such examples highlight how emotional tone changes as semantics involving multiple senses evolve.

However, most computational approaches treat semantic change and emotion analysis as distinct tasks. Semantic shift analysis often relies on diachronic comparisons of word embeddings [13; 26; 41] and token-level embeddings for sense disambiguation [12; 42], while emotion analysis uses static affective lexicons with emotional scores that are in some cases reconstructed via

Max Tiessler, Quim Motger, Florina Piroi, and Andreas Baumann. "EmoTracker—A New Framework for Modeling and Forecasting Diachronic Emotion Dynamics." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 760–784. https://doi.org/10.63744/tdBQckiQA3FI.

static word embeddings trained on historical text data [1; 5]. The reliance on word embeddings for this purpose, however, comes with certain limitations. Most prominently, word embeddings cannot straightforwardly differentiate between senses. For emotion analysis, this is evidently problematic since different senses of one and the same word can vary regarding their emotional connotation (e.g., 'awe-inspiring' vs. 'disgusting'). The main challenge in the current state-of-the-art lies in the lack of frameworks that model emotional evolution at a sense-specific level, bridging the relationship between polysemous meaning and emotion. Such a framework would be needed, however, to fully understand the emotional evolution of words [9].

To address this gap, we introduce EmoTracker, a novel framework that integrates diachronic sense modeling with temporal emotion analysis based on the Valence–Arousal–Dominance (VAD) model [36; 45] that captures three emotional dimensions (Valence: negative—positive; Arousal: calm—agitated; Dominance: submissive—dominant). Our contributions to computational humanities research are threefold: (1) we propose the first automatic method for constructing diachronic sense-informed VAD datasets by aligning historical sense distributions [18] with established contemporary VAD lexicons (NRC-VAD [33], Warriner [45], MEmoLon [6]); (2) we develop an LSTM-based forecasting model augmented with attention mechanisms and momentum-based features [29], which capture temporal changes in the emotional state, to predict future VAD trajectories, introducing a novel predictive capability for affective language evolution; and (3) we build a REST API and interactive interface that features 2D VAD and sense time series, 3D emotion-over-time plots and 4D visualizations that capture the diachronic joint evolution of emotions and senses in the VAD space.

We structure our work around the following research questions:

RQ1: How can we construct diachronic, sense-informed VAD datasets without manual annotation by systematically integrating temporal sense distributions with static contemporary VAD lexicons?

RQ2: How effectively can predictive models forecast future emotional shifts using enhanced LSTM-based architectures with temporal features?

RQ3: How can we design interactive visualizations to explore historical and forecasted emotional trajectories in multidimensional VAD space representations?

EmoTracker presents a reproducible and extensible framework that unifies diachronic semantics and affective modeling. We provide all data, open-source infrastructure and a complete replication package[1], facilitating adoption by computational humanities researchers to investigate how emotional meaning evolves with language over time. Upon publication, the data and code will be made available through an institutional research data repository.

2 Related Work

While diachronic sense modeling and emotion analysis have each developed robust methodologies, these fields are often treated independently. As a result, we lack frameworks that jointly capture how word senses and their emotional associations co-evolve over time. EmoTracker addresses this integration gap by linking temporal sense distributions with affective modeling, enabling new insights into diachronic language dynamics.

The past 15 years have seen considerable advances in the study of semantic change, initially driven by efficient ways of generating static embedding representations for large sets of words based on historical corpus data [26; 38; 41]. In these approaches, semantic change is often operationalized by measuring shifts in a word's embedding over time or that of its semantic neighbors

[1] Replication package: `https://github.com/mtiessler/EmoTracker`

[7; 13]. However, it has been proved that similarities between embeddings are sensitive with respect to frequency of occurrence so that this confound needs to be controlled for when studying semantic change [10]. Moreover, while such approaches implicitly capture the semantic neighborhood of a word, they typically lack detail about the evolution of a word's different senses. For this, word-sense disambiguation approaches are necessary. Such approaches often rely on clustering token-level embeddings [12; 42] or developing sense-disambiguation models that are trained on sense-annotated historical data [39].

In a related approach, Hu et al. [17] employ sentence embeddings derived with BERT from historical sense-specific example sentences that were taken from the Oxford English Dictionary to implement a pipeline that can disambiguate between word senses in historical English text data. Their pipeline was then used to predict a word's senses and derive their respective frequency of occurrence in the Corpus of Historical American English, spanning the 19th and 20th century. For each word and each period they thus provided a probability distribution over all its senses.

Their data were used to predict under what conditions words become more polysemous [2], and Kali et al. [23] use a similar pipeline to infer sense-specific data in order to examine predictors of word-sense decline. Making predictions, however, is disputed in historical linguistic research [37]. Language change is often seen as erratic and governed by a too complex set of interacting factors, to the effect that some scholars even argue that diachronic linguistics should not be a predictive science (see discussion in [37] and [44]). Still, predictive modeling techniques have been explored in the field [43]. Our approach contributes to this agenda by reconstructing and forecasting the emotional semantics of words, which is an important step for anticipating and studying shifts in public sentiment, detect emerging connotations, and support cultural and linguistic analysis over time.

Diachronic emotion analysis has applications in diverse domains such as literary studies [1; 28], but also economics [3] or public health [32]. In historical linguistics, research on lexical amelioration and pejoration is well established. Connected to this, Morin and Acerbi [34] revealed a decrease in positive terms in historical English texts, which is in line with pejoration cycles observed in lexical change, in which words tend to obtain more negative senses to the effect that they are replaced by more positive ones; cf. *toilet* vs. *bathroom* [15]. Interestingly, this is contrasted by the observation that speakers tend to use more positive than negative words, a phenomenon known as 'linguistic positivity bias´ [19].

To study phenomena like this, it is essential to have information about emotional status of words not only now but also for language stages several decades or centuries ago. In addition, such data should be ideally available for large sets of words. Cook and Stevenson [9] draw on PMI-based similarity of words to a pre-defined set of seed words in order to historically reconstruct lexical valence. Seed words, in this context, are words like *good* or *death* that are supposed to have had stable emotional semantics throughout the observation period. The reconstructed valence scores were then used to study pejoration and amelioration dynamics. Similarly, Fonteyn and Manjavacas [11] use embedding-based similarity together with a set of positive and negative seed words.

In a more general approach, Buechel et al. [5] use word embeddings to regress historical valence, arousal, and dominance scores (VAD) from seed words. The integration of the dimensions of arousal and dominance was particularly welcome given that dimensional approaches to modeling emotion have a long tradition [36] and that interactions between the three dimensions are well-known in cognitive research [16; 45]. Reconstructed scores were tested against a manually created gold-standard dataset.

While the requirement of creating a historically stable set of seed words could have been be relaxed [14], the reconstruction of emotion scores based on similarities among word embeddings is not without problems. For one, similarities between word embeddings can be affected by frequency [10], as discussed above. More severely, emotional properties of individual word senses cannot be

examined in this way. It was shown that annotators tend to only consider the prototypical, i.e., most common, sense in concreteness labeling tasks [35]. Applied to emotion analysis, this could mean that the emotional meaning of a less common sense is not well reflected in aggregated emotion scores, and such errors would be propagated through static word embeddings. In our approach, we infer sense-specific emotion scores from all sense descriptions associated with a word written in contemporary English [46] and in this way effectively circumvent the problems that come along with transferring emotional semantics via embedding similarities.

3 Methodology

We propose EmoTracker, a unified framework for modeling sense-informed emotional trajectories over time. Built on an adapted CRISP-DM process model [40] and illustrated in Figure 1, our methodology integrates diachronic sense modeling with temporal emotion tracking to capture how emotional meaning evolves across word senses.

The research method comprises six stages: (1) **automatic dataset construction**, aligning temporal word sense distributions with three VAD lexicons (NRC-VAD, Warriner, MEmoLon); (2) **dataset evaluation**, using a gold-standard diachronic emotion dataset [4] for quality assessment; (3) **neural model training**, employing an own designed LSTM-based architecture with momentum features and temporal attention [24]; (4) **model evaluation**, measuring predictive accuracy across VAD dimensions; (5) **API design**, exposing a REST interface for VAD trajectory forecasting; and (6) **interactive visualization**, enabling human-in-the-loop exploration of emotional and semantic change.

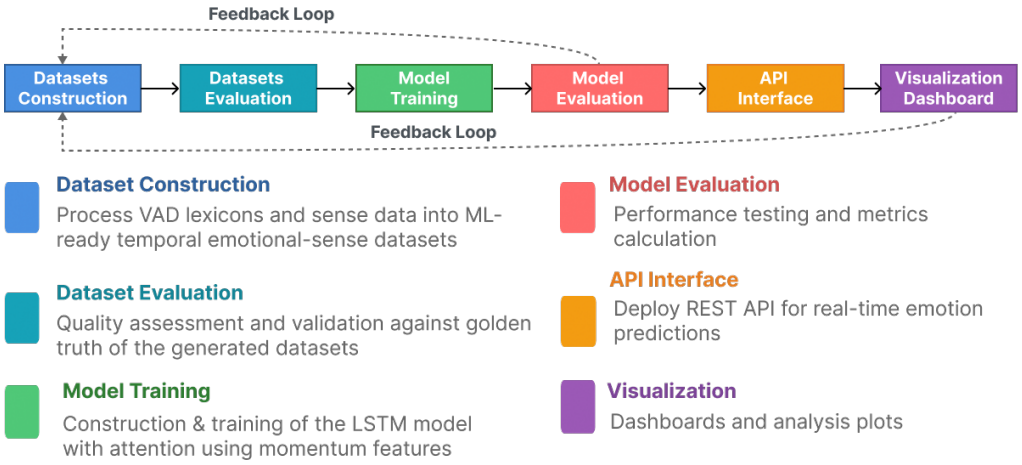


Figure 1: Workflow for developing the EmoTracker framework, showing the complete pipeline from dataset generation to deployment, with integrated feedback loops for iterative improvement.

A core strength of this methodology is its built-in flexibility. If an evaluation step results are unsatisfactory, users can re-execute the pipeline with alternative lexicons, allowing for iterative refinement and increased robustness of the constructed datasets.

One of the main outcomes of this design is a fully reproducible research package. EmoTracker includes all source code, datasets, trained models, and deployment configurations, enabling replication of the entire pipeline or individual stages. The full package is openly available via the EmoTracker repository.

3.1 Diachronic VAD Dataset Construction

At the time of writing to the best of our knowledge, no existing models were capable of performing VAD inference based on word-sense distributions, which required us to develop a new approach. However, training such a model requires appropriate data, which was also not available. While the current dataset landscape offers comprehensive diachronic sense data from 1830 to 2010, VAD lexicons remain temporally fragmented and limited to individual time points. This can be seen graphically in Figure 3 in Appendix C,

To address this gap, and given the lack of gold-standard temporal sense VAD datasets, we developed a novel method to construct large-scale diachronic sense VAD datasets. Our method consists of the following three reproducible steps:

1. **Data Integration:** Our datasets were created by integrating two main sources:
   (a) Diachronic sense proportions, $p(s_i, t)$, derived from a sense modeling dataset [18], where $p(s_i, t)$ denotes the proportion of sense s_i used at time t. The probability distributions over senses in this data set have originally been generated for 3,220 polysemous words by applying a sense-disambiguation model trained on sense-specific example sentences taken from the Oxford English Dictionary (OED) to a diachronic text corpus (Corpus of Historical American English) layered into decades (see [17] for details).
   (b) Three established VAD lexicons serving as sources for static 3-dimensional Valence Arousal Dominance (VAD) vectors: (1) The NRC-VAD lexicon [33], (2) The Warriner lexicon [45], (3) The MemoLon lexicon [6].
2. **Deriving Sense-Specific VAD ($VAD_{\text{sense}}(s_i)$):** For each individual word sense s_i, we computed a fixed 3-dimensional VAD vector. This process involved identifying the keywords surfacing in the chosen VAD lexicon within the sense's lexicographic OED-definition and retrieving their corresponding 3D VAD vectors from the respective lexicon [46]. Crucially, sense definitions are written in contemporary English, enabling the use of a contemporary VAD lexicon. The final vector for the sense, $VAD_{\text{sense}}(s_i)$, was derived by computing the element-wise average of these keyword vectors:
$$VAD_{\text{sense}}(s_i) = \frac{1}{|K_{s_i}|} \sum_{k \in K_{s_i}} VAD_{\text{keyword}}(k)$$
   where:
   - K_{s_i} is the set of keywords associated with the sense s_i,
   - $VAD_{\text{keyword}}(k)$ is the 3-dimensional VAD vector for keyword k from the lexicon.
3. **Calculating Diachronic Word VAD ($VAD_{\text{word}}(w, t)$):** To determine the final VAD score for a word w at a specific time t, we first define its set of constituent senses as $S_w = \{s_1, s_2, \ldots, s_n\}$, where n is the total number of senses for that word. The time-specific VAD score, denoted $VAD_{\text{word}}(w, t)$, is then calculated as a weighted average of the static VAD scores of its senses, using the sense probabilities as weights:
$$VAD_{\text{word}}(w, t) = \sum_{i=1}^{n} p(s_i, t) \cdot VAD_{\text{sense}}(s_i)$$
   where:

- $VAD_{\text{word}}(w, t)$ is the final, time-specific VAD vector for word w,
- $p(s_i, t)$ is the proportion of sense s_i for word w at time t,
- $VAD_{\text{sense}}(s_i)$ is the static 3-dimensional VAD vector for sense s_i,
- n is the number of senses for word w.

This three-step method was applied to each of the three diachronic VAD lexicons, resulting in separate diachronic VAD datasets, each containing reconstructed VAD values for all decades from 1820 to 2010 for all words in the intersection of [18] and the respective static VAD lexicon: (1) EmoTracker-NRC, (2) EmoTracker-Warriner, and (3) EmoTracker-MemoLon. Each was independently constructed using its respective lexicon as the source of static VAD values.

3.2 Datasets Evaluation

To assess the quality of our automatically constructed diachronic VAD datasets, we evaluated them against the GoldEN VAD dataset [4], a manually annotated historical gold standard from circa 1835. This benchmark provides expert-validated VAD scores for English words and is well-suited for temporal emotion analysis. The evaluation aimed to assess how closely our automatic VAD estimates align with expert-annotated historical values, evaluate the reliability of our methodology, and compare the performance of different lexicon sources in capturing historical emotional meaning. The process involved several key steps:

1. **Temporal Alignment:** For each of our three datasets, we extracted VAD estimates for the year 1835 to align with the gold standard temporal reference point.
2. **Word Matching:** We identified overlapping vocabulary between each constructed dataset and the gold standard, focusing on words present in both sources to ensure a proper comparison.
3. **Scale Normalization:** Given that our constructed datasets are on a [0,1] scale while the gold standard uses a [1,9] scale, we applied adaptive min-max scaling to normalize value ranges.
4. **Statistical Analysis:** We computed multiple evaluation metrics for each dataset:
    - Pearson correlation coefficients (r) for each VAD dimension and overall performance
    - Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for error quantification
    - Statistical significance testing (p-values) to assess correlation reliability

This evaluation allowed us to determine which lexicon source provided the most reliable emotional estimates in a diachronic setting.

3.3 Forecasting Model and Evaluation Framework

We frame the VAD trajectory prediction task as a time-series forecasting problem. The objective is to estimate how a word's emotional dimensions (i.e., Valence, Arousal, and Dominance (VAD)) change over time. We use diachronic VAD trajectories defined by the diachronic VAD scores generated in the previous steps for this purpose.

To model the long-term dependencies in emotional change, we use a Long Short-Term Memory (LSTM) neural network. LSTMs are well suited for this task because they can retain information across long input sequences, which is necessary to capture gradual changes in meaning and emotion [24].

Each time step in the sequence is represented by 27 features. These include the three VAD difference values (Δv, Δa, Δd) and 24 momentum-based features. The momentum features are computed using eight different metrics across each of the three VAD dimensions. These metrics come from financial time-series analysis and are adapted to treat VAD values like time-dependent prices [8].

Although inspired by finance, our temporal setup is different. The dataset covers 39 time points from 1820 to 2010, spaced at 5-year intervals. This resolution is achieved by interpolating a diachronic sense dataset that originally had 10-year intervals.

Momentum features are calculated using sliding windows of 5 to 10 steps, which correspond to 25 to 50 years. The model uses a lookback window of 15 steps (75 years) to predict VAD values 5 years into the future. This long input range is suitable for capturing the relatively slow pace of emotional change.

All parameters, including window sizes and time resolution, can be modified in a configuration file provided in the reproducibility package. The full list of momentum features is shown in Table 5 in Appendix B.

The architecture consists of a two-layer LSTM with 128 hidden units, followed by a multi-head attention mechanism with eight heads. This allows the model to focus on important parts of the input history when making predictions. The network also uses layer normalization, dropout for regularization [25], and GELU activation functions [27].

Training is performed for 100 epochs using the AdamW optimizer [30], along with regularization techniques to reduce overfitting.

Model performance is evaluated both quantitatively and qualitatively. For the quantitative evaluation, we use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). For the qualitative evaluation, we compare the forecasting behavior of models with and without momentum features. As shown in Appendix E, Figure 4, the model with momentum features produces smoother and more consistent forecasts that better reflect historical trends. In contrast, the version without these features tends to produce noisier and less reliable predictions.

3.4 API and Visualization (EmoTracker Dashboard)

For visualizing the 3D trajectories and interpreting the model forecasting, we developed an interactive frontend dashboard along with a lightweight REST API. The API provides a prediction endpoint that accepts word and time horizon parameters, returning forecasted VAD trajectories as JSON responses. The dashboard allows users to load any of the constructed datasets in the dataset construction step, select target words, and visually analyze both historical and forecasted VAD and Sense dynamics. The visualization features include (1) 2D time-series views for each individual sense and VAD dimension to evaluate sense emotional historical trends in a word (Figure 6 in Appendix F); (2) 2D VAD time series multi-word comparison plots to contrast emotional trends across different words (Figure 5 in Appendix F); (3) a novel 3D visualization of a word's trajectory through time in the VAD space (Figure 7 in Appendix F); (4) a 4D representation incorporating sense proportions as an additional visual dimension, where the fourth dimension is encoded through color intensity (Figure 8 in Appendix F), which allows for analyzing the impact that individual word senses have on the emotional trajectory; and (5) interactive controls for word selection and forecast horizon adjustment.

4 Results

4.1 Construction of Reliable Diachronic Sense-Informed VAD Datasets

We evaluated our automatic dataset construction method against the historical GoldEN VAD gold standard [4], using the 78 words shared between our constructed datasets and the gold standard as

the evaluation set. As shown in Table 1, performance varied across the three datasets. EmoTracker-NRC achieved the highest correlation with the gold standard (r = 0.287, $p < 0.001$), indicating the strongest alignment with human-annotated VAD values. While EmoTracker-Warriner slightly outperformed NRC in terms of error metrics (MAE and RMSE), we prioritize correlation as the primary evaluation measure, as it directly reflects the preservation of emotional ranking and trends. All correlations were statistically significant, suggesting that our automatic method effectively captures VAD patterns, even without human supervision and at large scale. However, it is important to note that this evaluation is limited to a single historical reference point (1835), and does not assess how well the model captures emotional trends over broader temporal ranges. This limitation constrains our possibility to generalize conclusions about long-term diachronic validity and remains as future work.

Dataset	**Pearson's r**	**p-value**	**MAE**	**RMSE**
EmoTracker-NRC	**0.287**	$\mathbf{7.98\times10^{-6}}$	1.363	1.703
EmoTracker-Warriner	0.274	2.13×10^{-5}	**1.317**	**1.655**
EmoTracker-MemoLon	0.179	0.006	1.600	1.931

Table 1: Evaluation results of constructed datasets against the GoldEN VAD historical gold standard. The best results per metric are shown in bold.

To provide a more fine-grained evaluation, we conducted a dimension-specific analysis, presented in Table 2. The results reveal variability in correlation across the VAD dimensions. EmoTracker-NRC achieved the highest correlations for Valence and Arousal, while Dominance proved more challenging. Notably, EmoTracker-Warriner achieved statistically significant correlations across all three dimensions and yielded the strongest performance for dominance. In contrast, EmoTracker-MemoLon demonstrated weaker performance overall, with only valence showing a significant correlation.

Dataset	**VAD Dimension**	**Pearson's r**	**p-value**
EmoTracker-NRC	**Valence**	**0.389**	4.39×10^{-4}
	Arousal	**0.339**	0.002
	Dominance	0.094	0.412
EmoTracker-Warriner	Valence	0.249	0.028
	Arousal	0.292	0.009
	Dominance	**0.280**	0.013
EmoTracker-MemoLon	Valence	0.224	0.048
	Arousal	0.172	0.131
	Dominance	0.156	0.173

Table 2: Dimension-wise correlation analysis across all EmoTracker datasets. The highest Pearson's r value for each VAD dimension is shown in bold.

These findings are further supported by the error metrics presented in Table 1, where EmoTracker-NRC exhibited the lowest MAE and RMSE, indicating stronger overall reliability. Taken together, the results suggest that our dataset construction method effectively enables automatic, large-scale, and sense-aware VAD labeling over time, thus addressing RQ1.

The final EmoTracker-NRC dataset comprises 2,935 unique words and 13,916 sense-level definitions, spanning 39 temporal steps from 1820 to 2010 in 5-year intervals, resulting in a total of

125,580 VAD entries. These words represent 5.36% coverage of the original NRC VAD lexicon (2,935 out of 54,801 entries). The dataset was split chronologically, with data up to 1980 used for training and data from 1985 onward reserved for validation. This split results into 52,830 training sequences (78.3%) and 14,675 test sequences (21.7%), covering all 2,935 unique words in both sets. This volume and temporal granularity ensure sufficient data diversity and historical depth to support effective LSTM training and generalization. Not only does the dataset offer enough sequence coverage for training and testing, but its temporal breadth also extends existing diachronic VAD resources.

4.2 Predictive Modeling of Emotional Trajectories

We evaluate the performance of our LSTM model in forecasting VAD trajectories across time. As shown in Table 3, the model achieves low mean absolute error (MAE) and root mean square error (RMSE), indicating strong predictive accuracy across the full vocabulary.

Metric	**Value**
MAE	0.013
RMSE	0.015
Training Loss	0.008
Validation Loss	0.012

Table 3: Overall performance metrics of the LSTM model on VAD trajectory forecasting.

Appendix G presents histograms of MAE and RAE values for all words in the EmoTracker-NRC dataset. These distributions are right-skewed, indicating that the model performs exceptionally well for the majority of words.

Table 4 reports how forecasting performance varies over different time horizons. As expected, prediction error increases with forecast distance. Nevertheless, the model maintains reasonable accuracy even at a 20-year horizon, which is suitable for tracking long-term diachronic semantic and emotional change.

Forecast Horizon	**MAE**	**RMSE**	**Pearson's r**	**p-value**
5 years	0.011	0.013	0.821	$<$0.001
10 years	0.016	0.021	0.734	$<$0.001
15 years	0.024	0.032	0.642	$<$0.001
20 years	0.035	0.047	0.531	$<$0.001

Table 4: LSTM forecasting performance across different time horizons. Correlation remains statistically significant even for long-term predictions.

Forecasting accuracy degrades gradually with longer time spans but retains sufficient predictive power for long-term applications.

A deeper analysis of the best and worst forecasted words is available in Appendix H. As expected, forecasting works best when emotion trajectories do not exhibit a lot of change. Performance is good, however, also for more complicated dynamics. For a qualitative perspective, we selected three representative words: *body* (stable emotional trajectory), *gay*, and *alien* (both exhibiting semantic shift). Table 6 in Appendix D compares the predicted and actual VAD values over five-year intervals.

The model shows near-perfect accuracy for stable emotional trajectories (*body*) and captures both direction and magnitude of semantic shifts for dynamic words (*gay*, *alien*). These results confirm that the EmoTracker LSTM model effectively captures diachronic VAD trends and thus addresses **RQ2**.

4.3 Visualizing Temporal Emotion Dynamics

To address **RQ3**, we designed and implemented an interactive visualization interface to explore diachronic emotional and semantic change. The EmoTracker Dashboard, introduced in Section 3.4, provides an interactive platform for exploring sense-informed VAD trajectories over time. Through a lightweight API and an intuitive user interface, the system allows users to load the EmoTracker constructed datasets, forecast trajectories, and explore the emotional evolution across different words and senses.

The dashboard supports a diverse set of visualizations: (1) individual and multi-sense 2D VAD time-series plots to track emotional and sense temporal patterns (Appendix Figures 6 and 5); (2) a novel 3D visualization of word trajectories through VAD space over time (Figure 7), which provides an intuitive spatial representation of how emotions evolve diachronically; (3) another novel 4D visualization that incorporates sense proportions as a dynamic fourth dimension, encoded through color intensity (Figure 8); and (4) interactive controls for word search, forecast horizon adjustment, and multi-word comparison.

These multidimensional visualizations offer a novel approach to emotion and sense modeling, allowing the exploration of complex VAD trajectories that are difficult to capture with traditional 2D plots or static time-point visualizations, while supporting deeper diachronic cultural and linguistic analysis.

5 Discussion

This paper presented a unified framework for modeling and forecasting the evolution of the emotional connotations of words over time. Our contributions are threefold. First, we introduced a method for constructing large-scale, diachronic, sense-informed VAD datasets by integrating temporal sense distributions with static affective lexicons. Second, we developed an LSTM-based forecasting model, augmented with momentum features and temporal attention, that effectively predicts long-term emotional trends. Third, we designed a visual analytics interface to explore and interpret emotion-sense trajectories through intuitive 2D, 3D, and 4D visualizations.

While our findings confirm that the EmoTracker pipeline produces meaningful and interpretable affective forecasts, several limitations must be highlighted. The most significant threat to validity is the scarcity of diachronic, human-annotated VAD datasets. Our evaluation relies only on the GoldEN VAD dataset from 1835, leaving the remainder of the historical timeline without direct human supervision. This raises the risk of overfitting to a single temporal benchmark and underscores the need for gold-standard annotations spanning multiple periods to ensure temporal validity. Although our quantitative evaluation shows statistically significant correlations with gold-standard values, the lack of multi-period validation limits claims of general accuracy across time. This limitation highlights the need for richer gold-standard datasets that span multiple time periods. EmoTracker is already designed to easily integrate such data, enabling retraining and fine-tuning without modifying the model or implementation. With future expert-annotated datasets across historical periods, the framework will more accurately capture genuine historical shifts in emotion and reduce patterns arising from modeling bias.

Additionally, the performance varied across the VAD dimensions. Dominance proved to be the most difficult to model, likely because of inconsistent lexicon coverage and conceptual ambiguity. Additionally, our approach may inherit biases from contemporary English definitions and static

lexicons that do not reflect historical affective norms, potentially introducing anachronistic emotion scores; further validation is needed to assess historical accuracy. Lexicon choice also had an impact on the results, with NRC-VAD showing stronger overall alignment but Warriner performing better on some error metrics. These differences highlight the importance of lexicon selection in diachronic emotion modeling. Finally, our framework evidently depends on the availability of sense-distribution data (here: [17]). Current efforts in collecting diachronic sense-annotations [39] are highly relevant in this regard. Such efforts are particularly important in order to extend diachronic emotion analysis to languages other than English and German [5].

Our contribution adds to the discussion about the predictive nature of language change [37; 43] by introducing state-of-the-art forecasting techniques for studying language change. Importantly, however, EmoTracker is not only useful for making predictions; it can also support hypothesis generation in historical and cultural studies. For example, the decline in valence for the word *gay* from the 1900s to the 2000s reflects both a change in meaning and broader social and cultural shifts. By linking changes in word meaning with emotional trends, EmoTracker opens new possibilities for research in cultural analytics, historical linguistics, and the digital humanities. Its forecasting feature can also help researchers identify potential future changes in culture and emotion by predicting how the emotional tone of words might evolve over time. Likewise, the framework can be adapted for the task of forecasting backwards in time or interpolation for periods for which insufficient historical data are available.

Finally, our interactive dashboard allows an exploratory analysis of word trajectories, making diachronic affective trends more interpretable and accessible. However, while promising, the usability and impact of these visualizations remain to be evaluated in applied humanities research settings.

6 Conclusions

We presented EmoTracker, a framework for modeling and forecasting the co-evolution of word meaning and emotional connotation over historical time. By integrating temporal sense modeling with affective lexicons and neural forecasting, EmoTracker enables scalable, sense-informed VAD analysis at a diachronic scale.

Our findings highlight the potential of this approach, but also underline key challenges, including limited gold-standard data and sense-specific resources. Despite these constraints, the framework provides a reproducible foundation for exploring emotional meaning evolution over time.

Looking ahead, future work includes exploring more advanced forecasting models, such as hierarchical LSTMs, Transformer-based architectures, and fine-tuned large language models (LLMs) for sense-level emotion prediction. Creating expert-annotated diachronic VAD datasets across time would strengthen evaluation and reduce reliance on static resources. The framework can also be extended to multilingual and domain-specific contexts, and its dashboard evaluated for interpretability in applied humanities research.

EmoTracker's approach to diachronic emotion evolution not only provides a more holistic framework for emotional and semantic analysis but also establishes a foundation for future interdisciplinary research between NLP, computational humanities, and cultural temporal analysis.

References

[1] Acerbi, Alberto, Lampos, Vasileios, Garnett, Philip, and Bentley, R Alexander. “The expression of emotions in 20th century books”. In: *PloS one 8*, no. 3 (2013), e59030.

[2] Baumann, Andreas, Stephan, Andreas, and Roth, Benjamin. “Seeing through the mess: evolutionary dynamics of lexical polysemy”. In: *Proceedings of the 2023 conference on empirical methods in natural language processing*. 2023, pp. 8745–8762.

[3] Bentley, R Alexander, Acerbi, Alberto, Ormerod, Paul, and Lampos, Vasileios. “Books average previous decade of economic misery”. In: *PloS one 9*, no. 1 (2014), e83147.

[4] Buechel, Sven and Hahn, Udo. “HistEmo: Historical Gold Emotion Lexicons”. `https://github.com/JULIELab/HistEmo/tree/master/historical_gold_lexicons`. 2017.

[5] Buechel, Sven, Hellrich, Johannes, and Hahn, Udo. “Feelings from the Past—Adapting affective lexicons for historical emotion analysis”. In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. 2016, pp. 54–61.

[6] Buechel, Sven, Rücker, Susanna, and Hahn, Udo. “Learning and Evaluating Emotion Lexicons for 91 Languages”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 1202–1217. DOI: `10.18653/v1/2020.acl-main.112`. URL: `https://aclanthology.org/2020.acl-main.112/`.

[7] Cassani, Giovanni, Bianchi, Federico, and Marelli, Marco. “Words with consistent diachronic usage patterns are learned earlier: A computational analysis using temporally aligned word embeddings”. In: *Cognitive science 45*, no. 4 (2021), e12963.

[8] Choi, Jaehyung. “Physical approach to price momentum and its application to momentum strategy”. In: *Physica A: Statistical Mechanics and its Applications 415* (Dec. 2014), pp. 61–72. ISSN: 0378-4371. DOI: `10.1016/j.physa.2014.07.075`. URL: `http://dx.doi.org/10.1016/j.physa.2014.07.075`.

[9] Cook, Paul and Stevenson, Suzanne. “Automatically Identifying Changes in the Semantic Orientation of Words.” In: *LREC*. 2010.

[10] Dubossarsky, Haim, Weinshall, Daphna, and Grossman, Eitan. “Outta control: Laws of semantic change and inherent biases in word representation models”. In: *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2017, pp. 1136–1145.

[11] Fonteyn, Lauren and Manjavacas, Enrique. “Adjusting Scope: A Computational Approach to Case-Driven Research on Semantic Change.” In: *CHR*. 2021, pp. 280–298.

[12] Giulianelli, Mario, Del Tredici, Marco, and Fernández, Raquel. “Analysing lexical semantic change with contextualised word representations”. In: *arXiv preprint arXiv:2004.14118* (2020).

[13] Hamilton, William L., Leskovec, Jure, and Jurafsky, Dan. “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1489–1501. DOI: `10.18653/v1/P16-1141`. URL: `https://aclanthology.org/P16-1141/`.

[14] Hellrich, Johannes, Buechel, Sven, and Hahn, Udo. “Modeling Word Emotion in Historical Language: Quantity Beats Supposed Stability in Seed Word Selection”. In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, ed. by Beatrice Alex, Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz. Minneapolis, USA: Association for Computational Linguistics, June 2019, pp. 1–11. DOI: 10.18653/v1/W19-2501. URL: https://aclanthology.org/W19-2501/.

[15] Hilpert, Martin. “Historical linguistics”. In: *Cognitive Linguistics: A Survey of Linguistic Subfields*, ed. by Ewa Dąbrowska and Dagmar Divjak. 2023, pp. 108–132.

[16] Hofmann, Markus J, Kuchinke, Lars, Tamm, Sascha, Võ, Melissa LH, and Jacobs, Arthur M. “Affective processing within 1/10th of a second: High arousal is necessary for early facilitative processing of negative but not positive words”. In: *Cognitive, Affective, & Behavioral Neuroscience 9*, no. 4 (2009), pp. 389–397.

[17] Hu, Renfen, Li, Shen, and Liang, Shichen. “Diachronic Sense Modeling with Deep Contextualized Word Embeddings”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 3899–3908.

[18] Hu, Ying, Kutuzov, Andrey, Velldal, Erik, and Øvrelid, Lilja. “Diachronic Word Sense Modeling”. https://github.com/iris2hu/diachronic-sense-modeling. 2021.

[19] Iliev, Rumen, Hoover, Joe, Dehghani, Morteza, and Axelrod, Robert. “Linguistic positivity in historical texts reflects dynamic environmental and psychological factors”. In: *Proceedings of the National Academy of Sciences 113*, no. 49 (2016), E7871–E7879.

[20] Investopedia Staff. “Chande Momentum Oscillator (CMO): Definition and Interpretation”. https://www.investopedia.com/terms/c/chandemomentumoscillator.asp. 2024.

[21] Investopedia Staff. “Relative Strength (RS): Definition and Use in Investing”. https://www.investopedia.com/terms/r/relativestrength.asp. 2024.

[22] Investopedia Staff. “What Is the Exponential Moving Average (EMA)?” https://www.investopedia.com/ask/answers/122314/what-exponential-moving-average-ema-formula-and-how-ema-calculated.asp. 2024.

[23] Kali, Aniket, Xu, Yang, and Stevenson, Suzanne. “Cognitive Factors in Word Sense Decline”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 46. 2024.

[24] Kong, Yaxuan, Wang, Zepu, Nie, Yuqi, Zhou, Tian, Zohren, Stefan, Liang, Yuxuan, Sun, Peng, and Wen, Qingsong. “Unlocking the Power of LSTM for Long Term Time Series Forecasting”. In: *arXiv preprint arXiv:2408.10006* (2024). Accepted at AAAI 2025. DOI: 10.48550/arXiv.2408.10006. URL: https://arxiv.org/abs/2408.10006.

[25] Kukačka, Jan, Golkov, Vladimir, and Cremers, Daniel. “Regularization for Deep Learning: A Taxonomy”. In: *arXiv preprint arXiv:1710.10686* (2017).

[26] Kutuzov, Andrey, Øvrelid, Lilja, Szymanski, Terrence, and Velldal, Erik. “Diachronic word embeddings and semantic shifts: a survey”. In: *Proceedings of the 27th International Conference on Computational Linguistics*, ed. by Emily M. Bender, Leon Derczynski, and Pierre Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1384–1397. URL: https://aclanthology.org/C18-1117/.

[27] Lee, Minhyeok. “GELU Activation Function in Deep Learning: A Comprehensive Mathematical Analysis and Performance”. In: *arXiv preprint arXiv:2305.12073* (2023).

[28] Leemans, Inger, Zwaan, Janneke M van der, Maks, Isa, Kuijpers, Erika, and Steenbergh, Kristine. “Mining Embodied Emotions: A Comparative Analysis of Sentiment and Emotion in Dutch Texts, 1600-1800.” In: *Digital Humanities Quarterly 11*, no. 4 (2017).

[29] Lim, Bryan and Zohren, Stefan. “Time-series forecasting with deep learning: a survey”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (2021). DOI: 10.1098/rsta.2020.0209. URL: http://dx.doi.org/10.1098/rsta.2020.0209.

[30] Loshchilov, Ilya and Hutter, Frank. “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations (ICLR)*. 2019. DOI: 10.48550/arXiv.1711.05101. eprint: 1711.05101. URL: https://arxiv.org/abs/1711.05101.

[31] MarketInOut. “Velocity Indicator – Definition and Interpretation”. https://www.marketinout.com/stock-screener/industry.php?picker=vacc. 2024.

[32] Metzler, Hannah, Rimé, Bernard, Pellert, Max, Niederkrotenthaler, Thomas, Di Natale, Anna, and Garcia, David. “Collective emotions during the COVID-19 outbreak.” In: *Emotion 23*, no. 3 (2023), p. 844.

[33] Mohammad, Saif M. “NRC Valence, Arousal, and Dominance (VAD) Lexicon”. https://saifmohammad.com/WebPages/nrc-vad.html. 2025.

[34] Morin, Olivier and Acerbi, Alberto. “Birth of the cool: a two-centuries decline in emotional expression in Anglophone fiction”. In: *Cognition and emotion 31*, no. 8 (2017), pp. 1663–1675.

[35] Reijnierse, W Gudrun, Burgers, Christian, Bolognesi, Marianna, and Krennmayr, Tina. “How polysemy affects concreteness ratings: The case of metaphor”. In: *Cognitive Science 43*, no. 8 (2019), e12779.

[36] Russell, James A. “A circumplex model of affect.” In: *Journal of personality and social psychology 39*, no. 6 (1980), p. 1161.

[37] Sanchez-Stockhammer, Christina. “Can we predict linguistic change? An introduction”. In: *Studies in Variation, Contacts and Change in English 16* (2015).

[38] Schlechtweg, Dominik, Hätty, Anna, Del Tredici, Marco, and Schulte im Walde, Sabine. “A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 732–746. DOI: 10.18653/v1/P19-1072. URL: https://aclanthology.org/P19-1072/.

[39] Schlechtweg, Dominik, Zamora-Reina, Frank D, Bravo-Marquez, Felipe, and Arefyev, Nikolay. “Sense through time: diachronic word sense annotations for word sense induction and Lexical Semantic Change Detection”. In: *Language Resources and Evaluation 59*, no. 2 (2025), pp. 1431–1465.

[40] Shimaoka, Andre Massahiro, Ferreira, Renato Cordeiro, and Goldman, Alfredo. “Leveraging XP and CRISP-DM for Agile Data Science Projects”. 2025. DOI: 10.48550/arXiv.2505.21603. URL: https://arxiv.org/abs/2505.21603.

[41] Tahmasebi, Nina, Borin, Lars, and Jatowt, Adam. “Survey of computational approaches to lexical semantic change detection”. In: *Computational approaches to semantic change 6*, no. 1 (2021).

[42] Tahmasebi, Nina and Dubossarsky, Haim. “Computational modeling of semantic change”. In: *arXiv preprint arXiv:2304.06337* (2023).

[43] Velde, Freek Van de and Keersmaekers, Alek. “What are the determinants of survival curves of words? An evolutionary linguistics approach”. In: *Evolutionary Linguistic Theory 2*, no. 2 (2020), pp. 127–137.

[44] Walkden, George. “Against mechanisms: Towards a minimal theory of change”. In: *Journal of Historical Syntax 5*, no. 32-39 (2021), pp. 1–27.

[45] Warriner, Amy B, Kuperman, Victor, and Brysbaert, Marc. “Norms of valence, arousal, and dominance for 13,915 English lemmas”. In: *Behavior Research Methods 45* (2013), pp. 1191–1207.

[46] Zad, Samira, Jimenez, Joshuan, and Finlayson, Mark. “Hell Hath No Fury? Correcting Bias in the NRC Emotion Lexicon”. In: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, ed. by Aida Mostafazadeh Davani, Douwe Kiela, Mathias Lambert, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem. Online: Association for Computational Linguistics, Aug. 2021, pp. 102–113. DOI: 10.18653/v1/2021.woah-1.11. URL: https://aclanthology.org/2021.woah-1.11/.

A LSTM model architecture

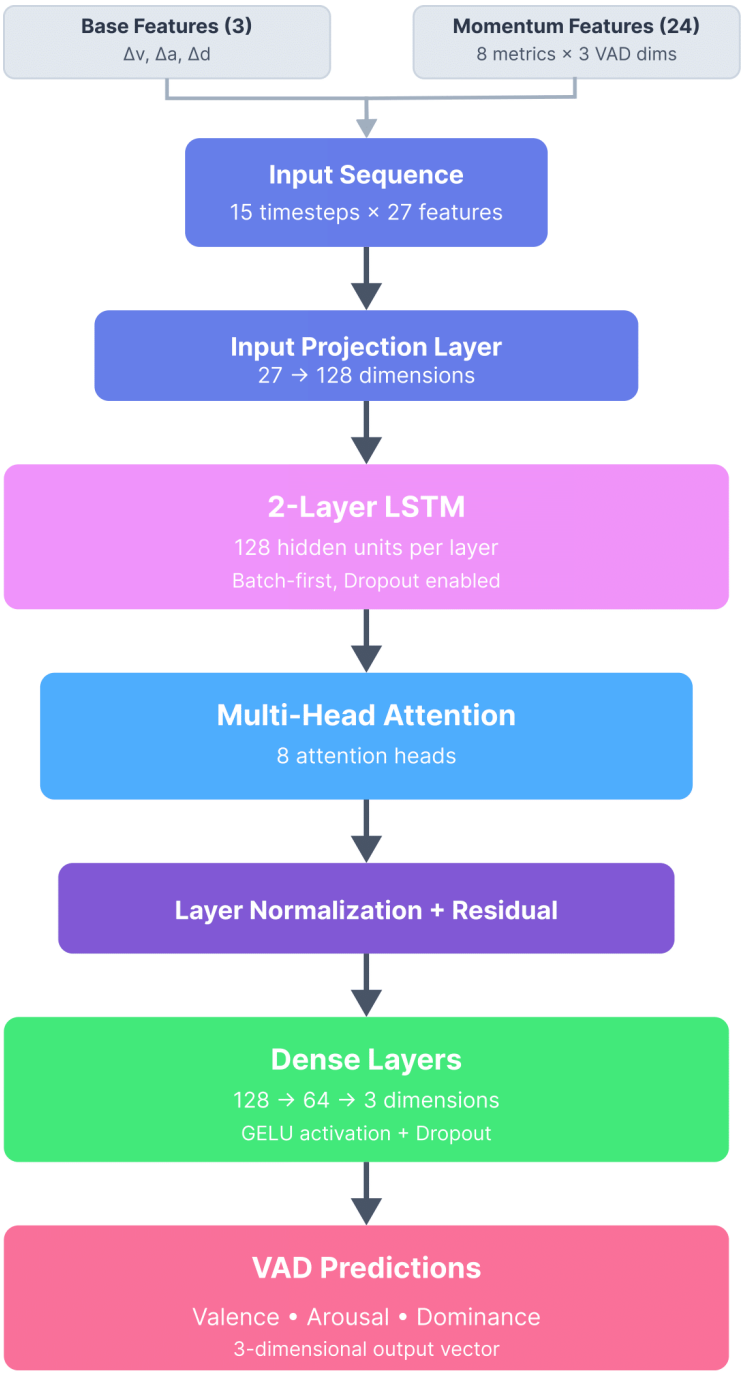


Figure 2: EmoTracker's LSTM architecture, combining momentum features, a 2-layer LSTM core, multi-head attention, and dense output layers for VAD prediction.

B Momentum Features Used in Forecasting

Feature	Description
Velocity [31]	The slope of a linear regression over the lookback window, indicating the trend's direction and speed.
Acceleration [31]	The second derivative, capturing the rate of change in velocity.
Trend Strength $\times$ **Direction**	The R-value from the linear regression, weighted by the trend's direction to measure consistency.
Volatility	The standard deviation of values in the window, measuring uncertainty and variability.
Momentum Oscillator [20]	The most recent change relative to the historical volatility within the window.
Relative Strength [21]	A comparison of the average value in the first half of the window versus the second half.
Range Position	The position of the current value relative to the historical minimum and maximum in the window.
EMA Ratio [22]	The ratio between a short-term Exponential Moving Average (EMA) and a longer-term Simple Moving Average (SMA) to identify trend crossovers.

Table 5: Description of Momentum Features Used in Forecasting

C Current Available Datasets

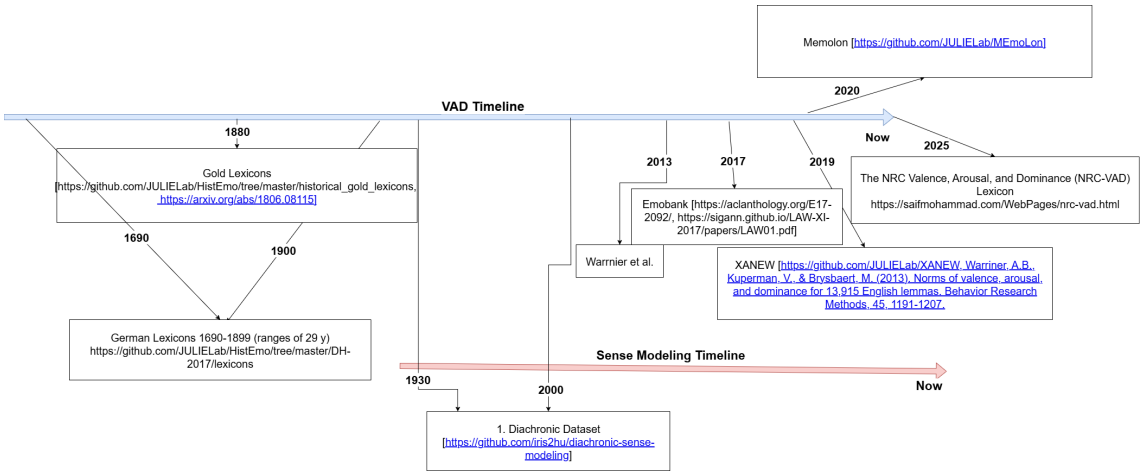


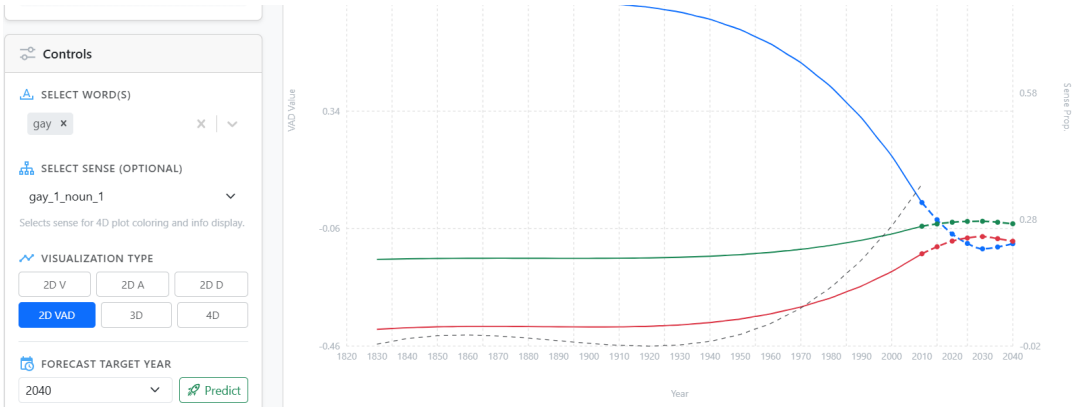
Figure 3: Integration of sense modeling and VAD lexicon resources (NRC-VAD, Warriner, MemoLon) in EmoTracker datasets.

D Predicted vs. Actual VAD Values

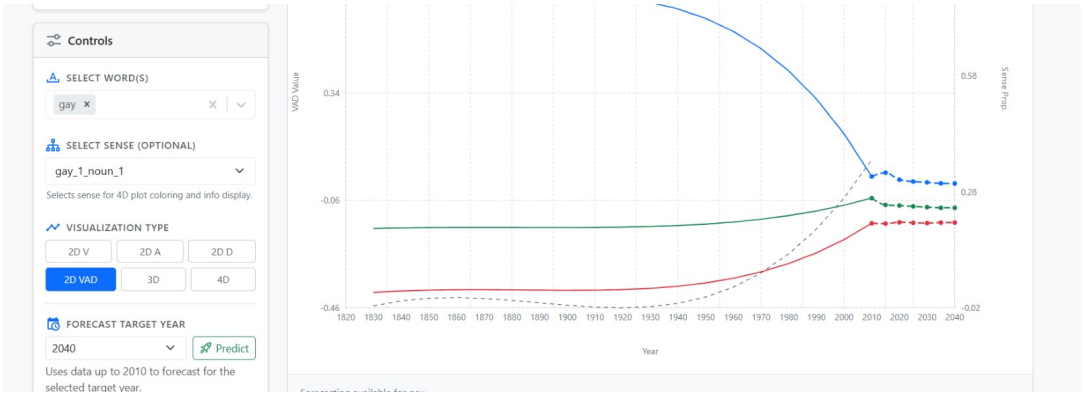
		Valence (V)		Arousal (A)		Dominance (D)	
Word	**Year**	**Predicted**	**Actual**	**Predicted**	**Actual**	**Predicted**	**Actual**
Stable trajectory word (minimal semantic shift)							
body	1990	0.1612	0.1612	−0.1780	−0.1779	0.0702	0.0703
body	1995	0.1623	0.1624	−0.1775	−0.1774	0.0728	0.0730
body	2000	0.1636	0.1636	−0.1771	−0.1770	0.0756	0.0757
body	2005	0.1654	0.1655	−0.1770	−0.1770	0.0790	0.0792
body	2010	0.1675	0.1675	−0.1771	−0.1770	0.0827	0.0828
Dynamic trajectory words (significant semantic shift)							
gay	1990	0.3233	0.3168	−0.2561	−0.2565	−0.1010	−0.1005
gay	1995	0.2669	0.2518	−0.2349	−0.2317	−0.0922	−0.0897
gay	2000	0.2009	0.1868	−0.2094	−0.2070	−0.0808	−0.0790
gay	2005	0.1337	0.1074	−0.1841	−0.1767	−0.0700	−0.0658
gay	2010	0.0539	0.0281	−0.1532	−0.1464	−0.0564	−0.0526
alien	1990	−0.1864	−0.1881	0.1286	0.1273	−0.1014	−0.1017
alien	1995	−0.1848	−0.1771	0.1202	0.1133	−0.0943	−0.0877
alien	2000	−0.1625	−0.1661	0.0977	0.0993	−0.0733	−0.0737
alien	2005	−0.1489	−0.1275	0.0853	0.0697	−0.0580	−0.0419
alien	2010	−0.0935	−0.0888	0.0447	0.0401	−0.0167	−0.0102

Table 6: Predicted vs. actual VAD values for representative words across time.

E Qualitative Comparison LSTM Model



(a) Forecast with momentum features



(b) Forecast without momentum features

Figure 4: Comparison of VAD trajectory forecasting for the word *"gay"* in 2040 using models (a) with and (b) without momentum features.

F EmoTracker Dashboard Views

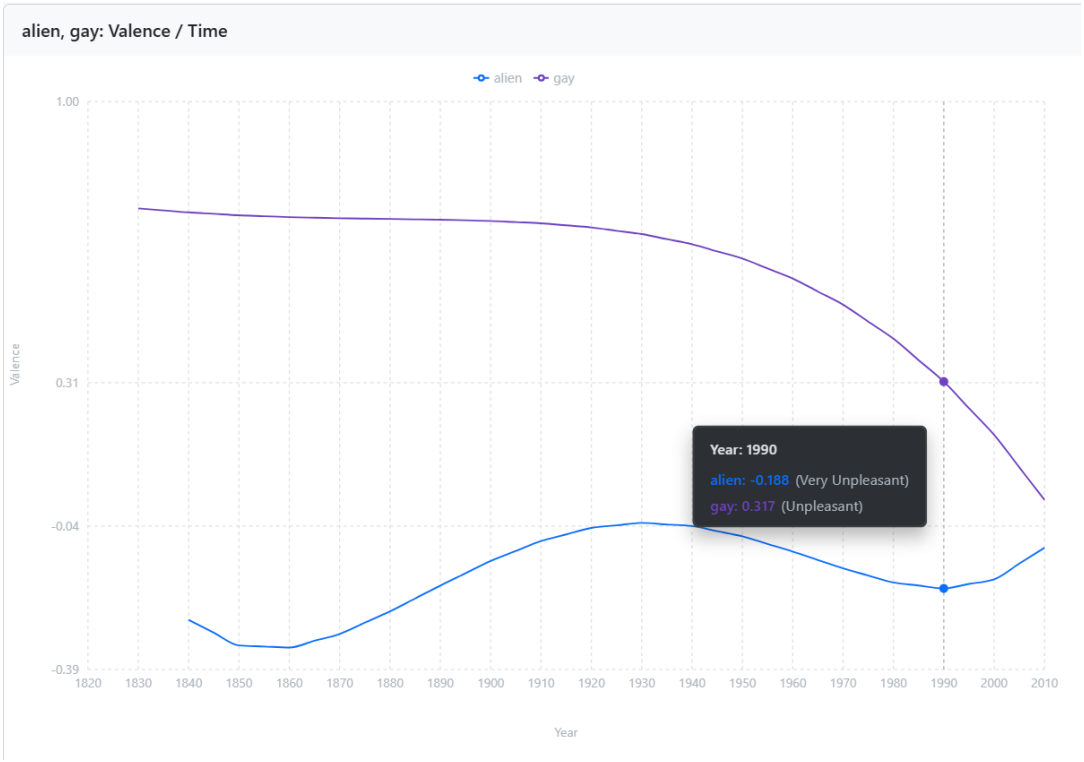


Figure 5: Valence trajectory comparison for *gay* and *alien*, showing historical shifts.

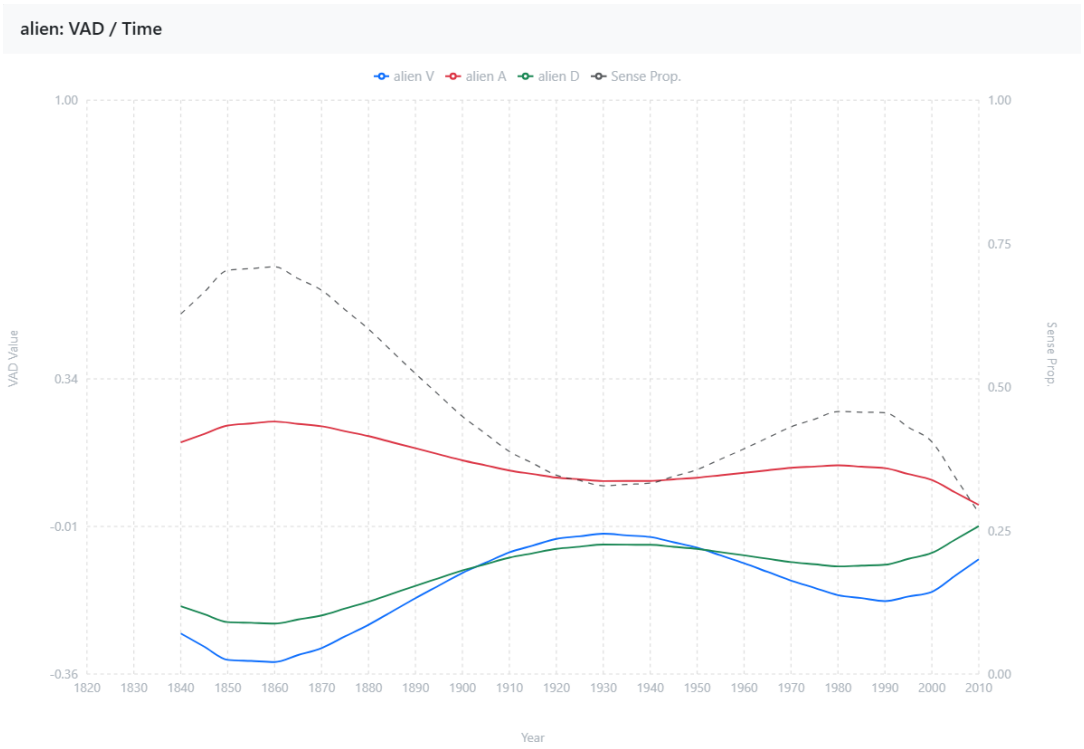


Figure 6: 2D time-series view of *alien*'s VAD (solid) and sense (dotted) evolution.

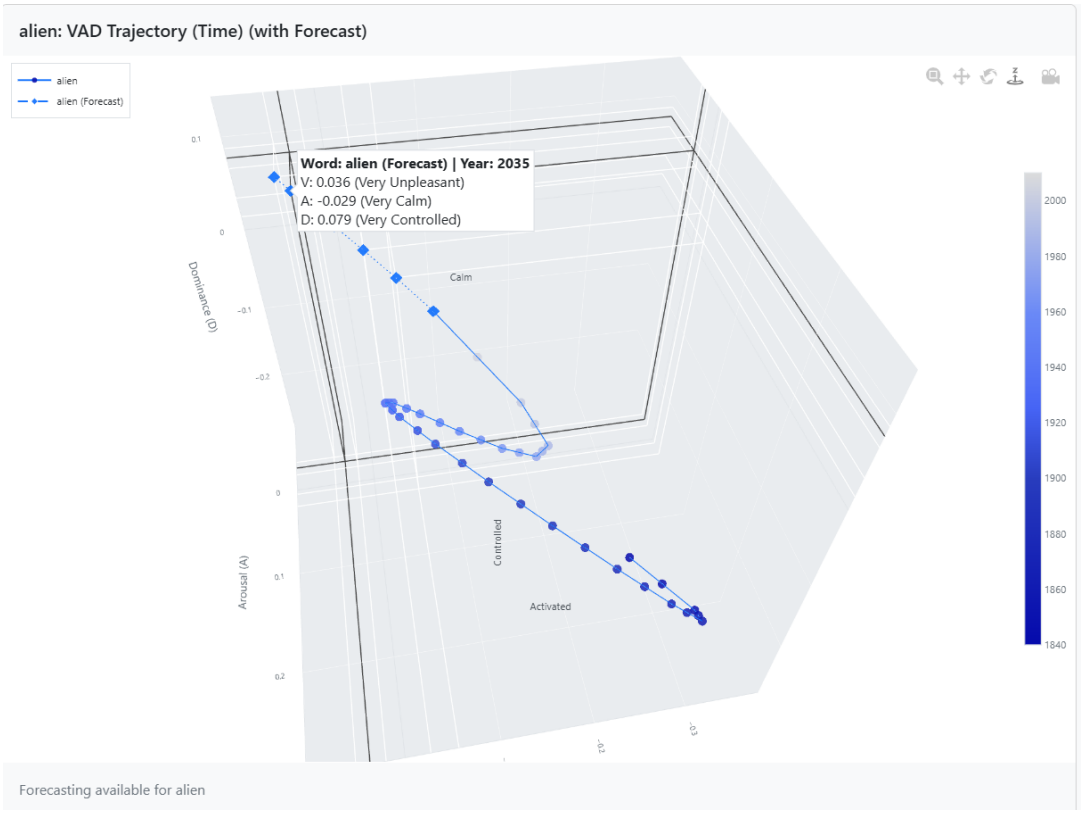


Figure 7: 3D visualization of *alien*'s VAD trajectory. Darker path = historical, lighter = forecasted.

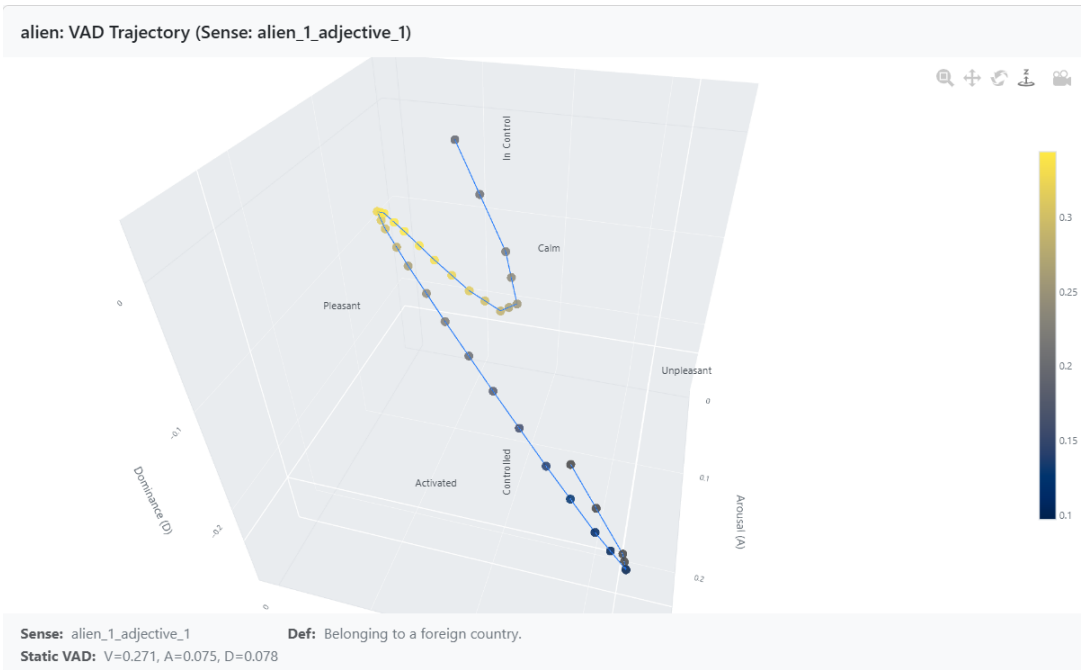


Figure 8: 4D visualization of emotion evolution. Color intensity encodes sense proportions.

G MAE and RMSE Distribution Analysis

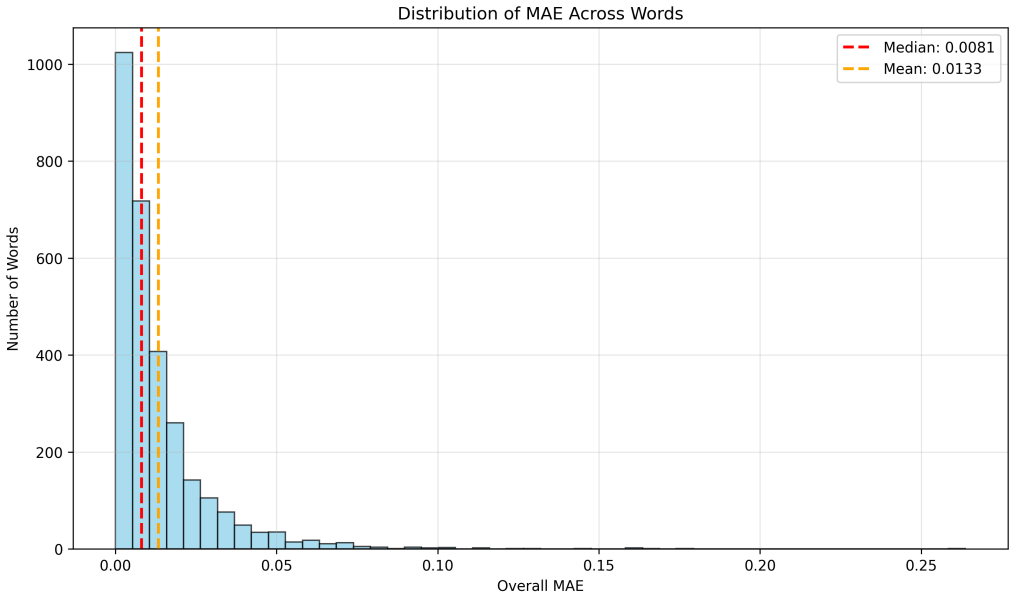


Figure 9: Distribution of MAE across all forecasted words. Median = 0.0081, Mean = 0.0133. The right-skewed shape indicates strong performance for most words.

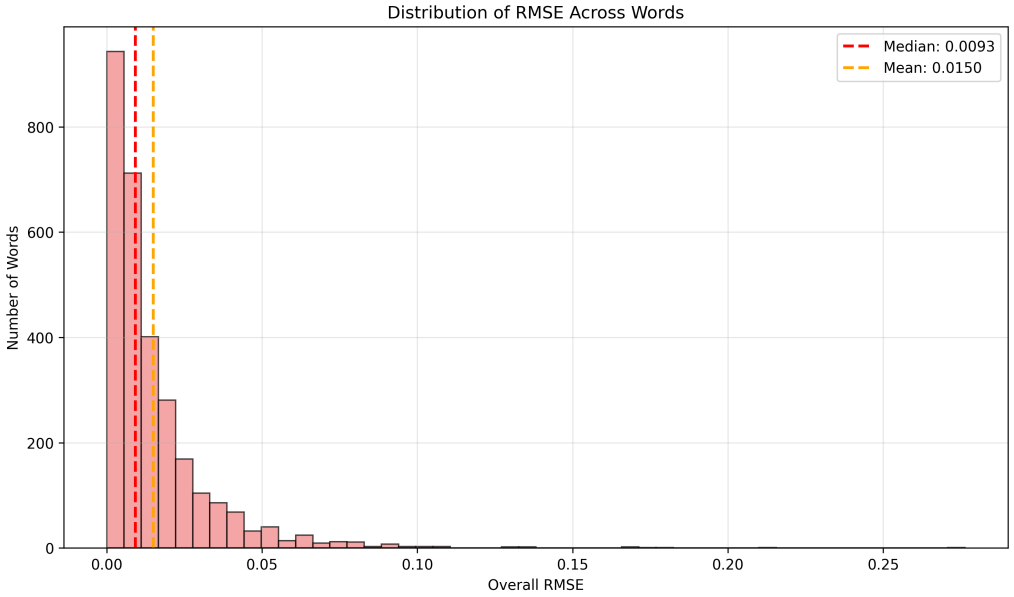


Figure 10: Distribution of RMSE across all forecasted words. Median = 0.0093, Mean = 0.0150. The right-skewed pattern supports the model's high accuracy for the majority of cases.

H Case Studies of Best- and Worst-Performing Words

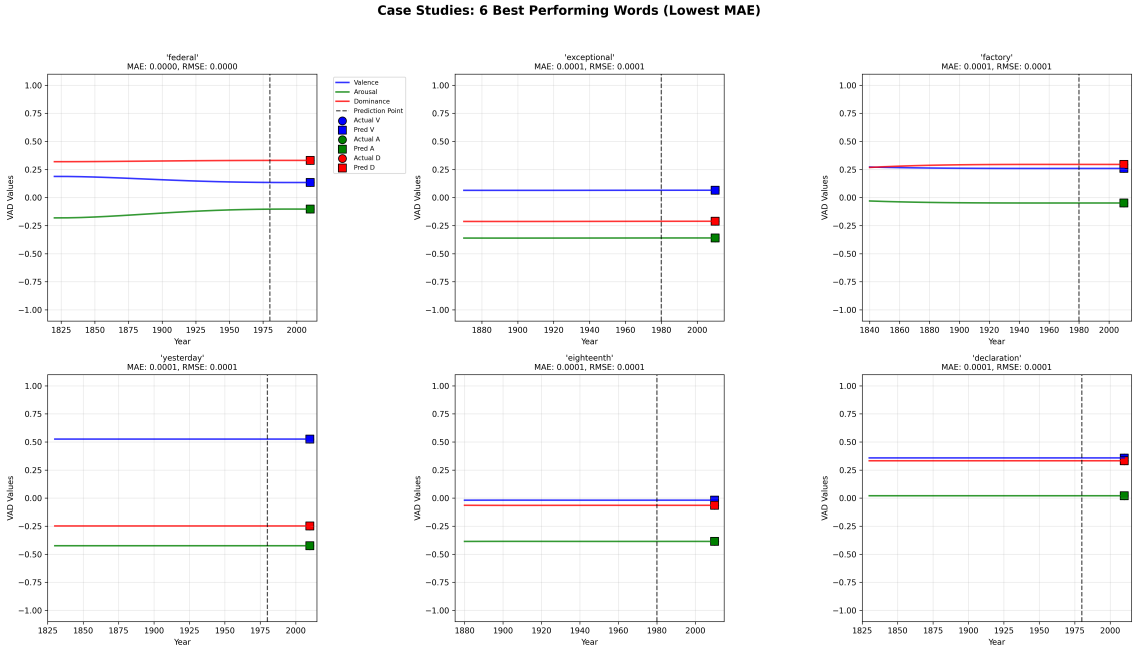


Figure 11: Top-performing words based on lowest MAE. These words exhibit high alignment between predicted and actual emotional trajectories.

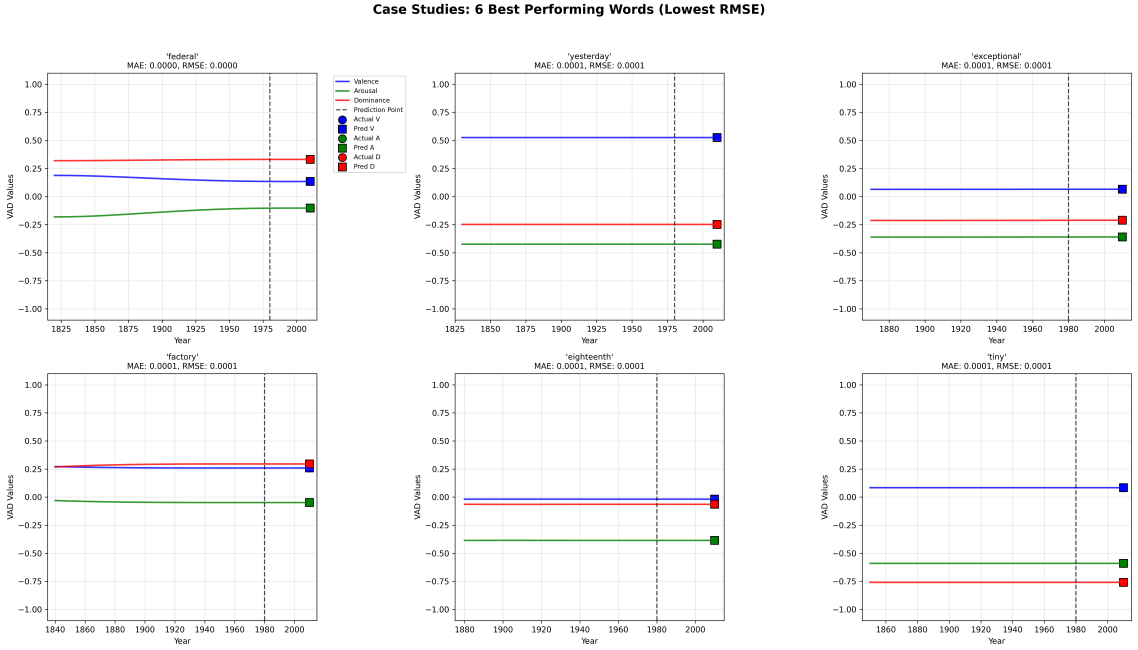


Figure 12: Top-performing words based on lowest RMSE. These words show highly stable and accurate predictions over time.

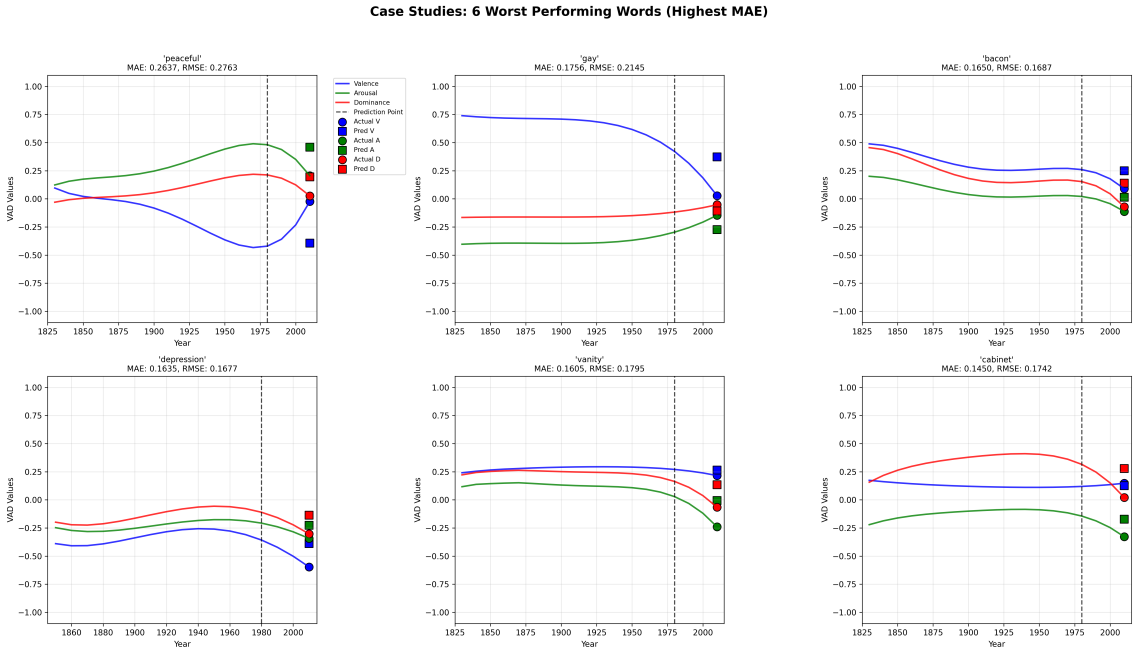


Figure 13: Worst-performing words by MAE. Errors suggest these words may exhibit irregular or complex semantic shifts.

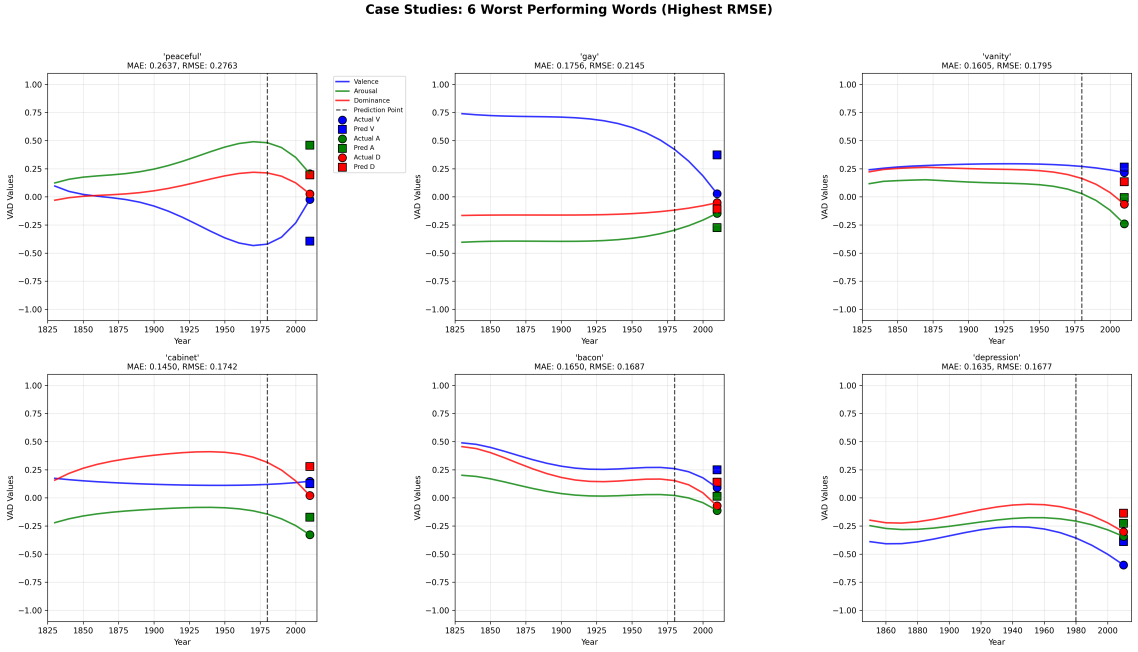


Figure 14: Worst-performing words by RMSE. Large prediction deviations may reflect ambiguous or noisy VAD histories.