Benchmarking Historical Phase Recognition from Text and Events

Fabio Celli[1,*], Marco Rovera[2]

[1]Maggioli Research, Santarcangelo di Romagna, Italy
[2]Fondazione Bruno Kessler, Trento, Italy

Abstract

This paper presents preliminary studies on a benchmark for the Historical Phase Recognition task. This task explores the application of computational linguistics to the study of long-term historical dynamics. We compare the utility of Event Tagging and BERT embeddings for classifying the phases of secular cycles defined by the the Structural-Demographic Theory. We explore this task both as five-class classification (crisis, growth, population immiseration, elite overproduction, State stess) and binary classification (rise, decline), on the basis of human- and LLM-annotated labels. Our findings reveal that Event Tagging, when aligned with human annotations, yields good performance in multi-class classification, but not in binary classification. Conversely, using BERT to extract features directly from text yields better performances with LLM-generated labels, in particular on the binary classification task. We also report higher inter-annotator agreement between LLMs compared to humans when labeling historical phases.

Keywords

Historical Phase Recognition, Cultural Analytics, Structural Demographic Theory, Large Language Models,

1. Introduction and Motivation

Historical Phase Recognition is a novel task that aims at the classification of phases of past societies according to existing theoretical frameworks. This task, based on the idea that history is a complex adaptive system [1] like language [2], can be useful for exploring and comparing societal adaptation processes in their long-term trends [3], to find replicable patterns. Societies have historical and structural dimensions [4] and evolve through dynamics that create cycles [5], following irreversible developmental paths that eventually cause them to break down [6] or recover. Crucially, much of historical information is expressed in natural language [7], and it is available from open sources like Wikipedia [8, 9], hence computational linguistics tasks such as event detection [10] can offer a great contribution to this line of research.

A theoretical framework in this area that has proven to be suitable for computational analysis is the Structural-Demographic Theory (SDT) [11]. By integrating this theory with data modeling techniques, researchers were able to make remarkably accurate predictions about the global crises that unfolded in the 2020s [12]. This predictive power underscores the value of SDT as a tool for analyzing complex socio-political dynamics within historical datasets [13]. Specifically, the SDT posits that historical cycles are characterized by five distinct phases:

- 0. Crisis (widespread conflict that results in a restructuring of the socio-political order);
- 1. Growth (a new order creates social cohesion, triggering high productivity and increasing competition for social status);
- 2. Population immiseration (increased competition for status and resources leads to rising inequality);
- 3. Elite overproduction (inequalities lead to radical factionalism and frustrated individuals who may become agents of instability) and
- 4. State stress (the rising instability brings fiscal distress and both lead the State towards potential crises with widespread conflicts, restarting the cycle).

SDT has proven to be a valuable framework for understanding a diverse array of historical occurrences. For instance, it has been applied to analyze the underlying causes of the French Revolution, the elite rivalries that fueled the American Civil War [14], and the factors contributing to the collapse of the Qing Dynasty [15]. Furthermore, SDT is also employed to analyze contemporary historical events, ranging from the Egyptian revolution of 2011 [16] to the political instability experienced in the US in 2021 [17].

Previous work in Historical Phase Recognition [18] released the Chronos dataset, annotated by humans, and demonstrated that systems can learn models with performance above chance, although far from perfect. Recent research in the field reports that LLMs can reach human performance in Historical Phase labeling and report that the intra-annotator agreement of LLMs is consistent [19].

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
fabio.celli@maggioli.it (F. Celli); m.rovera@fbk.eu (M. Rovera)
0000-0002-7309-5886 (F. Celli)

Still there is no benchmark in Historical Phase Recognition, and there are research questions about this task that remain unanswered, for instance:

- (RQ1) Can Event Tagging provide a generalization that helps Historical Phase Recognition?
- (RQ2) Can LLMs-as-annotators reach a higher consensus than humans in SDT labeling?
- (RQ3) Which kind of label is easier to model, the one made by humans or by LLMs?
- (RQ4) Is it easier to perform Historical Phase Recognition as 5-class or as a binary classification task?

To answer RQ1 we use EventNet-ITA, a Frame Parser[1] trained on a large Italian corpus, annotated with semantic frames of events[2]. This tool provides a fast and effective method for extracting Event Frames in Italian, achieving a performance of 0.9 F1-score for Frame Identification and 0.72 for Frame Element Identification on the original dataset [20]. To answer RQ2 we employ GPT4 [21] and Llama 3.1-400b [22] as annotators, producing a new SDT annotation on data. To answer RQ3 we adopt a perspectivist approach [23], running the classification task on different label sets and even on combination of labels. Lastly, to answer RQ4, we aggregate phases 1 and 2 under the label "rise" and phases 3, 4, and 0 under the label "decline," and then perform a binary classification task.

The paper is structured as follows: In Section 2 we describe how we created a benchmark from the Chronos dataset to promote the reproducibility of future experiments. In Section 3 we describe our experimental design, with annotation guidelines, prompts, analysis of labels and the results of the classification experiments. Finally, in Section 4, we draw our conclusion.

2. Data

Previous work on the Historical Phase Recognition task made a huge effort to produce annotated data [18], but the results of the previous classifications are not fully replicable. Hence we decided to develop a benchmark with fixed training and test sets out of the Chronos dataset.

The Chronos dataset, built upon the Seshat historical databank [24] and augmented with Wikipedia content, provides time-series data, in Italian and English, of historical events for 366 polities across 18 sampling zones, spanning from neolithic to the 2010s CE. Each row in the dataset represents an historical decade of a polity in a sampling zone. Textual descriptions of the selected events that happened in the decade include information about wars, reforms, rulers, population, elites, disasters, alliances, socio-economic context, famines, protests, elite changes, and religions. Descriptions are summarized to an average of 400 characters per decade, with source references when available. Each entry includes a timestamp, historical age, sampling zone, world region, and a standardized Polity ID encoding origin, name, societal type, and periodization. The dataset contains more than 9000 rows, but most of them have no textual description, especially those in remote times. Moreover, there are duplicates, as some polities expanded over more than one sampling zone, and were sampled more than once. The dataset also contains a flag to indicate whether the historical information reported is recorded or supposed. Using these information we created a benchmark.

2.1. Annotation and Agreement

First, we extracted event tags from the historical descriptions in Italian with EventNet-ITA. Then we removed duplicates and selected the rows with tags, text and recorded information. We obtained 1422 rows with data spanning from antiquity to 2010s. The data included also the original SDT labels, annotated by human hand following the points in these guidelines:

1. Read the textual description to identify key events: wars, reforms, rulers, population, elites, disasters, epidemics, alliances or treaties, socio-economic context, famines or financial stress, protests or movements, religions.
2. Use polity identifiers to find the start and end points of cultures. The end of a culture represents a crisis period.
3. Starting from the beginning of a culture, initially assign the sequence of labels of a standard secular cycle model: 1,1,2,2,3,3,4,4,4,0 and then evaluate whether to keep or change the labels in each decade. It is possible to have longer or shorter cycles. There can be only one label 0 (crisis) per cycle. A polity can have one or more cycles.
4. Having in mind the key events in the textual description, select one of the following labels to describe the decade: 1=growth. A society is generally poor when it experiences renewal or change followed by demographic (but not always territorial or economic) growth. Reforms, alliances, wars won or similar events are potential indicators of this phase. 2=impoverishment of the population. Potential economic and/or territorial expansion slows while demography continues to expand. The elite takes much of the wealth and defines the status symbols. Stability and external attacks are potential indicators of this phase. 3=Overproduction of the elites. The wealthy seek to translate their wealth into positions of authority and prestige. The population becomes poor.

[1] https://huggingface.co/mrovera/eventnet-ita
[2] https://huggingface.co/datasets/mrovera/eventnet-ita

Movements, protests, and wars are potential indicators of this phase. 4=State stress. The elites want to institutionalize their advantages in the form of low taxes and privileges that lead the state into fiscal difficulties. Wars, protests and changes in the elite are potential indicators of this phase. 0=Crisis. a triggering event such as a war, revolt, famine or disaster that the state is unable to manage leads to a new configuration of society. Emigration of elites, subjugation to other societies, civil wars or profound reforms are potential indicators of this phase.

5. Use the progressive order of the phases if no textual description is available for the decade.
6. Make sure there is a progressive order of the labels (e.g. phase 3 must follow phase 2). All labels can be repeated in the following decade except the crisis phase, which conventionally lasts one decade.

The annotation in the Chronos dataset was validated with three human annotators, who independently labeled a sample of 93 examples from the data. The initial agreement was low (Fleiss' k 0.206) because a single disagreement has an exponential impact on the rest of the sequence, but after a training session and the use of a standard pattern to start with (the sequence of secular cycle labels 1,1,2,2,3,3,4,4,4,0), the agreement between humans raised to Fleiss' k 0.455.

In order to answer RQ3 (whether it is easier to predict labels annotated by humans or LLMs) we produced new labels using GPT4 (1.8 trillion parameters) and Llama 3.1 (405 billion parameters) with the prompt reported in Figure 1 and temperature of 0.5. We provided the input data in chunks containing sequential decades of one or two polities per run. Despite the prompt explicitly required to assume that the sequence of labels follows a standard secular cycle model like the one used by humans (1,1,2,2,3,3,4,4,4,0), sometimes the LLMs produced as output unordered labels.

In order to create a benchmark, we split the data into training (1222 instances) and test set (200 instances). The labels have comparable distributions in the training and test set, as reported in Figure 2. While human and LLM labels approximate a log-normal distribution, the averaged labels approximate a normal distribution. This is because averaging labels with big misalignments (such as label "1" and label "4") tend to produce more labels "2", which became a wastebasket label.

We computed the inter annotator agreement over all 1422 examples and pairs of annotators, greatly expanding the experiments presented in literature. We evaluated results with k statistics and Krippendorf's α [25]. Although pairs that mix human and LLM annotations have an agreement comparable to previous results, here GPT4

Act as an expert historian and consider the Structural Demographic Theory (SDT). Given a set of descriptions of historical decades for different polities, label each description with one of the following secular cycle phases (sdtphase):
0=crisis (in this phase may happen societal collapse patterns, power transitions, conflicts, administrative or social structure changes, and external influences. Look for signs of civil wars, military coups, environmental factors, population movements, reform of tax systems, trade network disruptions, class conflicts, and foreign invasions). 1=growth (a society recovers from a crisis finding a new fresh culture that creates social cohesion. to recognize this phase examine the power structure patterns, legitimacy of rule, social organization, cultural elements, military aspects, and social changes. Look for the presence of strong elite classes, religious legitimation of power, centralized administrative systems, trade networks, cultural practices, territorial expansion, and population movements); 2=population impoverishment (growth slows and inequalities begin to emerge. to recognize this phase evaluate the power dynamics, economic patterns, military aspects, cultural/religious elements, administrative features, and infrastructure development. Look for succession struggles, trade route development, territorial conquests, religious tolerance, bureaucratic reforms, and construction projects); 3=elite overproduction (the number elite aspirants rises and the social lift mechanisms deteriorate. To recognize this phase assess power dynamics, governance, economic patterns, social structures, cultural and technological development, and common catalysts for change. Look for power struggles, trade system developments, social unrest between elite and population, religious developments, and military conflicts), 4=state stress (elites struggle to institutionalize their advantages. to recognize this phase review political instability, power struggles, economic challenges, military conflicts, administrative changes, and social/religious tensions. Look for succession disputes, financial crises, territorial loss, reforms to advantage specific elite groups, social unrest and religious conflicts). Initially assume that the sequence of labels follows a standard secular cycle model: 1,1,2,2,3,3,4,4,4,0 and then evaluate whether to keep or change the labels in each decade. Evaluate each label on the basis of the preceding and following ones. It is possible to have longer or shorter cycles. A cycle cannot turn back and cannot skip phases. So if in 1940 there is a phase 0, in 1950 there should be a phase 1, in 1960 there can be a phase 1 or phase 2. If in 1960 there is a phase 2, in 1970 there can be a phase 2 or phase 3, not a phase 4. If in 1970 there is a phase 3, in 1980 there can be a phase 3 or 4, and if in 2000 there is phase 4, in 2010 there can be a phase 0 or another phase 4. The decade after phase 0 the cycle restarts from phase 1.

This is an example of the input (json): $\langle example \rangle$
and this is the desired output (csv): $\langle example \rangle$
set of descriptions to label (json): $\langle data \rangle$

Figure 1: Prompt for the annotation of historical data with LLMs

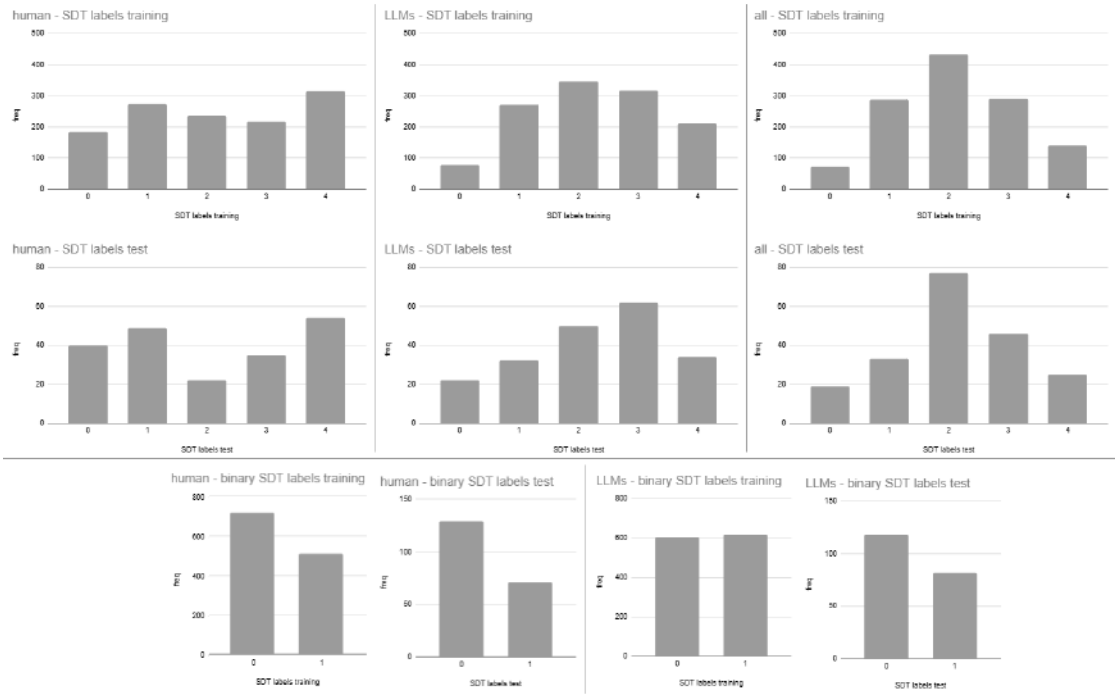

Figure 2: Distribution of labels for the multi-class classification task over the different label configurations.

and LLama3.1 have the highest score. This confirms, using a larger dataset, that LLMs can achieve a very high level of agreement on this task, even with temperature 0.5; moreover, these findings closely match the results obtained when both humans and LLMs received identical instructions and the temperature was set to zero [19]. The evaluation with Krippendorf's α, which could better capture the importance of label order, shows results similar to the ones computed with Fleiss and Cohen's k, suggesting that there might be disagreements on distant labels, like 0 and 4. Results are reported in Table 1.

Table 1
Results of inter-annotator agreement between pairs of Historical Phase annotators.

pair	Fleiss' k	Cohen's k	α
human+gpt4	0.215	0.218	0.216
human+llama3.1	0.211	0.212	0.211
llama3.1+gpt4	0.380	0.381	0.380

2.2. Contents

The final dataset contains the following features:

- a *decade ID* formatted with a standard method: 2 letters to indicate the area of origin of the culture, 3 letters to indicate the name of the polity, 1 letter to indicate the type of society (c=culture/community; n=nomads; e=empire; k=kingdom; r=republic), 1 letter to indicate the periodization (t=terminal; l=late; m=middle; e=early; f=formative; i=initial; *=any) and a number corresponding to the decade. For example "EgPdyk*-2960" is the pre-dynastic kingdom of Egypt in the 2960s b.C. "ItRomrm-220" is the middle Roman Republic in the 220s b.C. and "TrOttet1850" is the terminal phase of the Ottoman Empire in the 1850s;
- a *short Italian textual description* of the decade (the one used for the experiments);
- a *short English textual description* of the decade;
- the list of *tags* extracted from text;
- *human annotated SDT labels*;
- SDT labels annotated with *GPT4*,
- SDT labels annotated with *Llama3.1*,
- the *average of all the SDT labels*, turned into integer values;
- the *average of the SDT labels generated with LLMs*, turned into integer values;
- the *binary labels annotated by humans* obtained from SDT labels (1,2=rise; 3,4,0=decline);
- the *binary labels annotated by LLMs* obtained from SDT labels (1,2=rise; 3,4,0=decline).

Examples of data follows[3]:

1. JpKamk*1290, “al tempo del reggente Hōjō Sadatoki (r. 1284–1301) per il principe Hisaaki il clan Hōjō era alleato del clan Adachi. Tuttavia un complotto di Adachi Yasumori per usurpare gli Hōjō portò al colpo di stato noto come incidente Shimotsuki. vinse Hojo.”,“at the time of Regent Hōjō Sadatoki (r. 1284–1301) for Prince Hisaaki the Hōjō clan was allied of the Adachi clan. However a plot by Adachi Yasumori to usurp the Hōjō resulted in the coup known as Shimotsuki incident. the Hōjō won.”, PROCESS*PROCESS_START ACTIVISTS*POLITICAL_ACTIONS INVADER*INVADING PROCESS_START POLITICAL_ACTIONS INVADING,4,4,4,4,4,0,0
2. IqBabke-1750, “possibile apertura di una rotta commerciale per beni di lusso e minerale di stagno verso il Levante (Caanan) e l’Anatolia orientale (occupata dagli Assiri).”,“possible opening of a commercial route for luxury goods and tin ore towards the Levant (Caanan) and eastern Anatolia (occupied by Assyrians).”, LAND*OCCUPANCY OCCUPIER*OCCUPANCY OCCUPANCY,2,2,2,2,2,1,1
3. EgMamke1340,“peste nera ad Alessandria nel 1347. Serie di sultani di breve durata.”,“black death in Alexandria in 1347. Series of short lived Sultans.”, OLD*TAKE_PLACE_OF KILLER*KILLING CAUSE*DEATH PLACE*DEATH TIME*DEATH TAKE_PLACE_OF KILLING DEATH,4,1,3,3,2,0,1

Example 1 describes the Japanese Kamakura period in 1290s and is a case where all the annotations agree about phase 4 (or 0, "decline" in the case of binary labels). Example 2 reports a description of Kassite Babylon in 1750s b.C. and is a case where all annotations agree on phase 2 (or 1, "rise"). Example 3 describes Mamluk Egypt in 1340s and it is a case of disagreement between annotations.

We ordered the data alphabetically using the text column, thus obtaining a pseudo-randomization of the instances and breaking the temporal sequences. We dubbed this dataset "Chronos benchmark", which is freely available on Huggingface[4].

3. Analysis and Discussion

In order to answer RQ1 (whether Event Tagging is useful to recognize different phases), we performed an analysis of events per label. To do so, we extracted wordclouds including only the examples where all annotators agreed on the same label. Figure 3 reports the wordclouds for the binary classification task. As introduced in Section 2, Event Frames are shown in uppercase, while Frame Elements in small caps, along with their Frame, in the format FRAME_ELEMENT*EVENT_FRAME. The larger and bolder a word, the more strongly it is associated with that particular phase. From the wordclouds is clear that there are overlapping Event Frames between the two phases (eg: CONQUERING, WAR, CHANGE_OF_LEADERSHIP, BEAT_OPPONENT), while the same Frame Elements seem to have different frequencies in the two phases.

Figure 3: Wordclouds of Event tags in the binary classification task. The wordclouds include only the examples where all annotations agreed on the same label. Event frames are represented in uppercase while frame elements in lowercase.

Things are much more complicated in the multi-class classification task, depicted in Figure 4. In summary, the wordclouds show a progression where there are many overlaps of Event Frames between phases, in particular the BEAT_OPPONENT and CONQUERING events. However, Frame Elements help distinguish between phases: THEME*CONQUERING clearly appears in the growth and crisis phases, while other low-frequency elements, such as PROCESS*PROCESS_START, and GOAL*ATTEMPT are distinctive of phases 3 and 4 respectively. In general, wordclouds with smaller words, like the ones for phase 2, 3 and 4, highlight the need to capture weak signals for the classification tasks.

Overall, the similarity of the tags between phases illustrate well how difficult is the Historical Phase Recognition task.

[3] EVENT_FRAMES are shown in uppercase, FRAME_ELEMENTS in small caps.

[4] https://huggingface.co/datasets/facells/chronos-historical-sdt-benchmark

Figure 4: Wordclouds of Event tags in the multiclass classification task. The wordclouds include only the examples where all annotations agreed on the same label. Event frames are represented in uppercase while frame elements in lowercase.

3.1. Experiments

In order to answer the research questions listed in Section 1, we performed two distinct tasks: a multi-class classification, and a binary classification. Both tasks have comparable settings, with 768 features extracted with a frequency token matrix from the EventNet-ITA tags (events) and 768 features extracted with BERT-Italian-XXL (bert). To ensure replicability, we used Learnipy [26], a suite of algorithms for data science and machine learning in Colab Notebooks available online[5],

Table 2 reports the balanced accuracy of different classification models: Naive Bayes (nb), Gradient Boosting (xgb), Linear Discriminant Analysis (lda) using the two feature extraction methods (events, bert) to predict the 5 SDT phases. The models were trained and evaluated on different sets of labels: human-annotated (human), an average of LLM annotations (llms), and an average of all annotations (all). The baseline for this task is 0.2.

Table 2
Results of the 5-class classification task. We used two feature extraction techniques, EventNet-ITA (events) and BERT-Italian-XXL (bert), with three classification algorithms, Naive Bayes (nb), Gradient Boosting (xgb), Linear Discriminant Analisys (lda) to classify the labels provided in the Chronos dataset (human), averaged between GPT4 and Llama3.1 (llms), and averaged over all the preceding labels (all). The metric is Balanced Accuracy, the baseline is 0.2. The best averaged value for each pair are marked in bold, the ones below the baseline are marked in italics.

labels	features	nb	xgb	lda	avg
human	events	0.249	0.250	0.212	0.237
human	bert	0.178	0.180	0.218	*0.191*
llms	events	0.234	0.191	0.227	0.217
llms	bert	0.197	0.213	0.248	0.219
all	events	0.251	0.250	0.237	**0.246**
all	bert	0.205	0.208	0.118	*0.176*

Interestingly, the combination of human labels, event tags and an algorithm that captures weak signals (Gradient Boosting) yields good performances, suggesting that for the 5-class classification the event-based features align well with the human understanding of the SDT phases. However, the more robust results are achieved using event tags on the average of all labels, possibly for the normal distribution resulted from averaging the labels. In contrast, BERT struggles with human labels: the results show an average balanced accuracy lower than the baseline.

This might indicate that the contextual embeddings from BERT, while powerful, don't directly capture the nuances of the SDT phases as effectively as the event-based

[5] https://colab.research.google.com/drive/1G1VNHUCoDTso6wIWmrdvM21Z6D1PC6nL?usp=sharing

features when aligned with human annotations. However, the best performance when using labels averaged from LLMs is achieved with BERT features and Linear Discriminant Analysis. This hints that the patterns captured by BERT might be more consistent with the way LLMs interpret and label the SDT phases, although less transparent.

An interesting point is that event tags show consistent performance across different label sets (human, all, llms). The event tagger features consistently provide competitive results, often outperforming or closely matching BERT, with the advantage of being transparent. This highlights the value of explicit event information for this Historical Phase Recognition task. Overall, performance still needs improvement. While some results surpass the baseline of 0.2, the balanced accuracy scores indicate that accurately classifying the 5 SDT phases remains a challenging task.

Table 3 presents the results of the binary classification task, where the 5 SDT phases were aggregated into "rise" (phases 1 and 2) and "decline" (phases 0, 3, and 4). The same feature extraction methods and classification algorithms were used on human-derived binary labels (human) and LLM-averaged binary labels (llms).

Table 3
Results of the binary classification task. We used two feature extraction techniques, EventNet-ITA (events) and BERT-Italian-XXL (bert), with three classification algorithms, Naive Bayes (nb), Gradient Boosting (xgb), Linear Discriminant Analisys (lda) to classify binary labels computed from the Chronos dataset (human2), and averaged between the ones annotated by GPT4 and LLama3.1-400b (llms2). The metric is Balanced Accuracy, the baseline is 0.5, and the best averaged value is marked in bold, the ones below the baseline are marked in italics.

labels	features	nb	xgb	lda	avg
human	events	0.510	0.504	0.471	*0,494*
human	bert	0.489	0.477	0.558	0,508
llms	events	0.541	0.512	0.507	0,52
llms	bert	0.509	0.534	0.553	**0,532**

In this case, when combining BERT with LLM-averaged binary labels, we obtain a good average balanced accuracy. This confirms that BERT embeddings are particularly well-suited for capturing the broader temporal trends as interpreted by the LLMs.

In general, the performance with the binary labels is better with LLM annotations, implying that LLM-as-annotators are a promising technique for binary task in Historical Phase Recognition, also because their inter-annotator-agreement is generally better than the one reached by humans.

4. Conclusion

In conclusion, this study has taken initial steps in leveraging computational linguistics for the complex task of Historical Phase Recognition within the Structural-Demographic Theory framework. Our investigation into the utility of Event Tagging revealed its promise, particularly when aligned with human-annotated data, achieving the most robust performance in the 5-class classification task. This suggests that explicitly identified event structures resonate with human understanding of SDT's nuanced phases. Conversely, while powerful, BERT embeddings struggled to capture these nuances as effectively on human labels, hinting at a potential mismatch between its learned representations and the human interpretation of SDT.

Interestingly, BERT showed better performance with LLM-generated labels, indicating a possible alignment in their interpretation patterns, albeit with a loss of transparency compared to event tags. Answering RQ1 (whether Event tagging is useful): our results show that event tags help Historical Phase Recognition when coupled with human annotations. Instead, having LLM-generated labels, transformer models seem the best choice. In general our results show similar improvements over the baseline with the multi-class and binary classification tasks. Hence, answering RQ3 (which kind of label is easier to model), we can say there is no big difference. However, answering RQ4 (which classification task is easier), our results suggest that makes more sense to perform Historical Phase Recognition either as 5-class task with human annotated label and event tags, or as binary classification with LLM-annotated labels and BERT. Looking ahead, further research should explore methods to enhance the representational power of both event-based features and contextual embeddings for this task. Investigating techniques to better align LLM interpretations with human understanding of historical theories, and exploring more sophisticated classification models.

Our results also show that LLMs-as-annotators reach a higher consensus than humans in SDT labeling, and this answers RQ2. Since historical annotation is costly, time consuming and prone to bias, it is more likely that in the future we will see more LLM-annotated data. This suggests that the most promising future direction is having Historical Phase Recognition as a binary classification tasks. Ultimately, the integration of computational linguistics with historical theory holds significant potential for advancing our capability of extracting long-term societal dynamics from unstructured sources, and enhance our understanding of the cyclical patterns that shape human history. Given the general poor performance in Historical phase Recognition, we suggest there is still great room for improvement.

Author Contributions Ctatement

F.C.: conceptualization, experiments and main manuscript text; M.R.: data enrichment with Event Tagging, manuscript editing. All authors edited and reviewed the manuscript.

Acknowledgments

This research was supported by the European Commission, grant 101120657: European Lighthouse to Manifest Trustworthy and Green AI—ENFIELD.

References

[1] C. E. Maldonado, History as an increasingly complex system, History and Cultural Identity: Retrieving the Past, Shaping the Future (2011) 129–152.

[2] K. Lund, P. Basso Fossali, A. Mazur, M. Ollagnier-Beldame, Language is a complex adaptive system: Explorations and evidence, Language Science Press, 2022.

[3] A. Toynbee's, A study of history, Munich: List. Henry, William P., Greek Historical Writing: A Historiographical Essay (1991).

[4] N. Luhmann, D. Baecker, P. Gilgen, Introduction to systems theory, Polity Cambridge, 2013.

[5] R. Dalio, Principles for dealing with the changing world order: Why nations succeed or fail, Simon and Schuster, 2021.

[6] I. Wallerstein, Historical systems as complex systems, European Journal of Operational Research 30 (1987) 203–207.

[7] K. Lai, J. R. Porter, M. Amodeo, D. Miller, M. Marston, S. Armal, A natural language processing approach to understanding context in the extraction and geocoding of historical floods, storms, and adaptation measures, Information Processing & Management 59 (2022) 102735.

[8] M. Fisichella, A. Ceroni, Event detection in wikipedia edit history improved by documents web based automatic assessment, Big Data and Cognitive Computing 5 (2021) 34.

[9] M. Rovera, A knowledge-based framework for events representation and reuse from historical archives, in: European Semantic Web Conference, Springer, 2016, pp. 845–852.

[10] R. Sprugnoli, S. Tonelli, One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective, Natural language engineering 23 (2017) 485–506.

[11] J. A. Goldstone, Demographic structural theory: 25 years on, Cliodynamics 8 (2017).

[12] P. Turchin, Political instability may be a contributor in the coming decade, Nature 463 (2010) 608–608.

[13] P. Turchin, A. Korotayev, The 2010 structural-demographic forecast for the 2010–2020 decade: A retrospective assessment, PloS one 15 (2020).

[14] P. Turchin, A Structural-Demographic Analysis of American History, Beresta Books Chaplin, 2016.

[15] G. Orlandi, D. Hoyer, H. Zhao, J. S. Bennett, M. Benam, K. Kohn, P. Turchin, Structural-demographic analysis of the qing dynasty (1644–1912) collapse in china, Plos one 18 (2023) e0289748.

[16] A. Korotayev, J. Zinkina, Egypt's 2011 revolution: A demographic structural analysis, in: Handbook of revolutions in the 21st century: The new waves of revolutions, and the causes and effects of disruptive political change, Springer, 2022, pp. 651–683.

[17] P. Turchin, End times: elites, counter-elites, and the path of political disintegration, Penguin, 2023.

[18] F. Celli, V. Basile, History repeats: Historical phase recognition from short texts, Proceedings of CLIC-it 2024 (2024).

[19] F. Celli, V. Basile, Large language models rival human performance in historical labeling, in: Proceedings of ARDUOUS 2025, co-located with ECAI, 2025.

[20] M. Rovera, Eventnet-ita: Italian frame parsing for events, in: Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), 2024, pp. 77–90.

[21] J. A. Baktash, M. Dawodi, Gpt-4: A review on advancements and opportunities in natural language processing, arXiv preprint arXiv:2305.03195 (2023).

[22] A. Deroy, S. Maity, Code generation and algorithmic problem solving using llama 3.1 405b, arXiv preprint arXiv:2409.19027 (2024).

[23] S. Frenda, G. Abercrombie, V. Basile, A. Pedrani, R. Panizzon, A. T. Cignarella, C. Marco, D. Bernardi, Perspectivist approaches to natural language processing: a survey, Language Resources and Evaluation (2024) 1–28.

[24] P. Turchin, H. Whitehouse, P. François, D. Hoyer, A. Alves, J. Baines, D. Baker, M. Bartokiak, J. Bates, J. Bennet, et al., An introduction to seshat: Global history databank, Journal of Cognitive Historiography 5 (2020) 115–123.

[25] K. Krippendorff, Computing krippendorff's alpha-reliability (2011).

[26] F. Celli, C. Casadei, Learnipy: a Repository for Teaching Machine Learning Without Coding, Technical Report, 2022. URL: https://github.com/facells/fabio-celli-publications/blob/main/docs/2022_learnipy_techreport.pdf.