When Trump Shakes Biden's Hand: Tracking Real-World Interactions in Time and Space from Textual Records

Anonymous ACL submission

Abstract

Interpersonal interactions reveal critical insights across political, cultural, and historical domains. Existing research predominantly analyzes online interactions (e.g., retweets, mentions, follows), but these virtual signals poorly capture real-world encounters (e.g., collaborations, negotiations, debates) and suffer from biases (fake accounts, bots, unrepresentative samples) while spanning only recent decades. In contrast, **real-world interactions in historical texts (encyclopedias, news, books) capture actual encounters across centuries with temporal and spatial contexts**. However, extracting structured interaction quadruplets *(Person1, Person2, Time, Location)* is challenging due to complex spatio-temporal dependencies across scattered and indirect mentions. **We construct the *WikiInteraction* dataset (4,507 annotated quadruplets) and propose FALCON**, integrating AR-BERT, multi-task learning, and feature transfer to model these dependencies. FALCON achieves 86.51% F1, outperforming all baselines including GPT-o1, while incurring significantly lower computational costs than LLMs, and demonstrates cross-domain applicability on *Encyclopedia Britannica* and news corpora. Applied at scale, it extracts **658,966 Wikipedia quadruplets over 1,000 years and 77,836 news quadruplets**—the largest spatio-temporal interaction dataset to date. A case study on US political polarization (1960–2024, 7,891 figures) reveals long-term geo-temporal trends invisible in online data. Code and datasets are available[1].

1 Introduction

Interpersonal interactions, especially among notable individuals, reveal insights into cultural, economic, political, scientific, and historical perspectives (Jackson, 2011; O'Neill, 2014; Cruz et al., 2017; Fuller and Wang, 2021), as seen in scientists (Newman, 2001; Fortunato et al., 2018), politicians (Hsu and Park, 2012; Plotkowiak and Stanoevska-Slabeva, 2013), and authors (Börner et al., 2004; Sun et al., 2011). Much existing research focuses on social media interactions (e.g., retweets, mentions, follows) (Kleinberg, 2013; Kubin and Von Sikorski, 2021; Donkers and Ziegler, 2025), offering scale and accessibility. However, these virtual signals poorly capture real-world encounters (e.g., collaborations, negotiations) and suffer fundamental limitations. In political polarization, uneven platform adoption leads to skewed datasets (Pokhriyal et al., 2023), while fake news (Van der Linden et al., 2020), bots (Deb et al., 2019), and coordinated disinformation (Nizzoli et al., 2021; Cinus et al., 2025) compromise analysis. Moreover, social media only covers recent decades, limiting historical analyses.

In contrast, textual records like encyclopedias, news, and books provide an ideal source for extracting long-term interactions with temporal and spatial contexts. While LLMs demonstrate strong extraction capabilities, corpus-level mining remains prohibitively expensive—processing Wikipedia with GPT-o1 exceeds $1.2 million (Appendix A.5). Task-specific models can match LLM performance at significantly lower cost (Chen and Varoquaux, 2025; Obeidat et al., 2025), making specialized approaches essential for large-scale spatio-temporal interaction extraction.

However, existing task-specific methods (Gergaud et al., 2016; Tao and Zhang, 2020) focus solely on person pairs, omitting time and location—two dimensions fundamental to understanding social dynamics (Barabâsi et al., 2002; Crandall et al., 2010) **but challenging to extract jointly**.

We formalize this new task: extracting interaction quadruplets *(Person1, Person2, Time, Location)* from unstructured text, where *Person1* and *Person2* engage in a direct interaction at *Time* in

[1] https://anonymous.4open.science/r/FALCON-7EF9. Full datasets will be available upon acceptance.

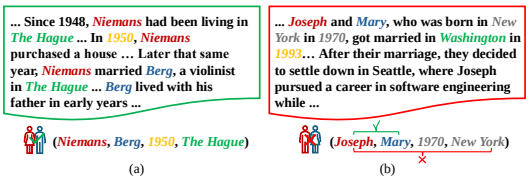


Figure 1: Example of extracted quadruplets and their contexts. (a) A valid spatio-temporal interaction where scattered entities must be linked. (b) An incorrect interaction inferred from mere proximity.

Location. Following prior work (Nijila and Kala, 2018; Labatut and Bost, 2019), we define an interaction as a direct action between two individuals (e.g., marriage, collaboration, competition).

Prior work (Tang et al., 2007; Gergaud et al., 2016; Tao and Zhang, 2020) identifies interactions by classifying person pairs, but extending to quadruplets faces two challenges. **First, entity association is unreliable**: heuristic methods applying NER to detect nearby temporal and spatial entities often fail—the nearest entities may not correspond to the actual interaction. In Fig. 1(b), this incorrectly identifies (*Joseph*, *Mary*, *1970*, *New York*), whereas they met in *Washington* in *1993*. Similarly, spatio-temporal co-occurrence methods (Zhang et al., 2025) produce false positives. **Second, mentions are scattered with varying importance**: relevant entities are dispersed across text and connected through indirect mentions. In Fig. 1(a), *Niemans* is first mentioned in *The Hague*, then associated with *1950*, and finally linked to *Berg*. Accurately linking these scattered mentions requires careful aggregation, as not all mentions are equally important.

Our Approach. To address these challenges, we propose FALCON (AR-Bert model utilizing Feature Transfer and Multi-Task Learning strategies for extracting spatio-temporal Life InteraCtiONs). FALCON reframes the problem as selecting correct quadruplets from all possible NER combinations. Given candidate quadruplets (*Person1*, *Person2*, *Time*, *Location*) with their contexts, FALCON classifies them as "interaction" or "not interaction."

We decompose this into two complementary tasks: (1) a **main task** determining whether *Person1* and *Person2* interact at a given *Time* and *Location*, and (2) an **auxiliary task** defined as a "trajectory task," which verifies their spatio-temporal co-occurrence by classifying triplets (*Person*, *Time*, *Location*) as valid trajectories. The auxiliary task provides reliable spatio-temporal evidence that directly supports the main task. This multi-task learning framework encourages the model to capture spatio-temporal correlations crucial for both tasks, while transfer learning incorporates features from the auxiliary task into interaction detection. To handle scattered mentions with varying importance, we utilize AR-Bert (Attention-enhanced R-Bert) as our feature extractor. By augmenting R-Bert (Wu and He, 2019) with an attention mechanism, it dynamically assigns higher weights to informative occurrences, overcoming the limitation of treating all mentions equally.

We select Wikipedia biographies as our primary extraction target and construct *WikiInteraction*, a manually annotated dataset with 4,507 quadruplets. **Trained on Wikipedia, FALCON applies well to Encyclopedia Britannica and news archives.**

Applied at scale, FALCON extracts **658,966 Wikipedia quadruplets** spanning over 1,000 years and **77,836 news quadruplets**. A case study on US political polarization (1960-2024, 7,891 figures) constructs a weighted network of adversarial, cooperative, and neutral interactions. Unlike studies relying on polls (Fiorina and Abrams, 2008; Pew Research Center, 2017, 2019), social media (Conover et al., 2011; Garimella and Weber, 2017; Flamino et al., 2023), or behavioral analysis (Fos et al., 2022), our approach reveals long-term geo-temporal trends invisible in online data, uncovering escalating polarization since the 1960s, accelerating sharply under Obama and Trump.

Our contributions include:

- We introduce the task of extracting spatio-temporal interactions from unstructured text and construct *WikiInteraction*, a manually annotated benchmark dataset.
- We propose **FALCON**, integrating multi-task learning, transfer learning, and AR-BERT. It achieves 86.51% F1, outperforming both LLMs and specialized models, while incurring significantly lower computational costs than LLMs. Furthermore, it demonstrates superior cross-domain applicability.
- We release 658,966 Wikipedia quadruplets spanning over 1,000 years and 77,836 news quadruplets, constituting the largest spatio-temporal interaction dataset to date, along with code to facilitate future research. A 60-year political polarization case study reveals geo-temporal trends invisible in online data.

2 Related Work

2.1 Social Media vs. Real-World Interactions

Social media data has been extensively used to study phenomena such as political polarization (Conover et al., 2011), scientific collaboration (Romero et al., 2011), and cultural diffusion (Centola, 2010). However, these studies face critical limitations: restricted temporal scope (Tufekci, 2014), noise from bots and disinformation (Allcott and Gentzkow, 2017; Cinus et al., 2025), and representational bias (Lutz and Hoffmann, 2017). In contrast, real-world interaction data derived from historical archives offers authentic interpersonal dynamics spanning centuries, as demonstrated in studies of 15th-century Florentine families (Padgett and Ansell, 1993) or Broadway collaborations (Uzzi and Spiro, 2005). Unfortunately, such studies are typically confined to small-scale, domain-specific cases. Our work bridges this gap by providing large-scale, structured real-world interaction data with precise spatio-temporal attributes, enabling broader research on network evolution (Barabâsi et al., 2002) and geographic dynamics (Onnela et al., 2011).

2.2 Interaction Extraction from Text

Extracting interpersonal interactions has evolved from early rule-based methods (Backstrom et al., 2006; Tang et al., 2007) to neural approaches. Researchers initially employed Named Entity Recognition with co-occurrence rules (Gergaud et al., 2016) or trigger word detection (Agarwal, 2016). Deep learning models like CNNs (Nijila and Kala, 2018) and Bi-LSTMs (Tao and Zhang, 2020) further improved performance by capturing semantic context. Recently, Large Language Models have shown promise in general information extraction (Xu et al., 2024). However, applying LLMs to corpus-level mining remains computationally prohibitive. Crucially, existing methods primarily focus on binary interaction detection, ignoring *when* and *where* interactions occur. We address this by proposing a novel task to extract spatio-temporal interaction quadruplets, utilizing a specialized framework incorporating multi-task learning and attention mechanisms.

3 Formulation of Task

Before we formulate the task, we first obtain candidate quadruples (*Person1*, *Person2*, *Time*, *Location*) using a heuristic method and attach the paragraph from which they originate. Our task is to determine, for each candidate quadruple (*Person1*, *Person2*, *Time*, *Location*), whether the two individuals interact at the specified time and location within the corresponding paragraph. We define "interaction" as the scenario where two individuals appear at the same time and place, with some form of connection between them. Such connections include direct conversations, collaborative activities, participation in the same event, or affiliation.

Given quadruple q and its text segment s, we train a model f with parameters Θ to classify q into $y \in \{0, 1\}$, where $y = 1$ indicates interaction between *Person1* and *Person2* at the specified *Time* and *Location*, and $y = 0$ indicates no interaction:

$$f : \{s, q, \Theta\} \rightarrow y$$

The heuristic method to obtain candidate quadruples is described in the Dataset section.

4 Dataset

Since there is no available dataset that captures the focus of our study, spatio-temporal interactions, we annotate a new ground truth dataset. The following section details the process of creating the $WikiInteraction$ dataset.

4.1 Data Collection

The dataset is derived from the biography sections of the English Wikipedia. We source the list of individuals and the links to their respective Wikipedia biography pages through Wikidata (Möller et al., 2022). In all, we have 1,930,519 people and their biography pages

4.2 Data Processing and Labeling

This section details the methodology used to extract and label potential quadruples in the format of (*Person1*, *Person2*, *Time*, *Location*) from the biography pages.

4.2.1 Extracting Quadruples

To initialize our extraction process, we leverage a trajectory mining pipeline adapted from (Zhang et al., 2024). This method combines NER to detect entity mentions with dependency syntax analysis to verify grammatical associations, thereby extracting individual spatio-temporal states in the form of triplets $(\textit{Person}, \textit{Time}, \textit{Location})$. We then construct candidate interaction quadruples $(\textit{Person1}, \textit{Person2}, \textit{Time}, \textit{Location})$ by align-

ing these trajectories based on strict spatio-temporal co-occurrence—specifically, pairing distinct individuals who are mapped to the identical time and location context within the text. Through this procedure, we obtained a total of 4,226,049 candidate interaction quadruples.

To evaluate the coverage of our method, we compare the interactions mentioned in the original pages with those extracted from the target text segment. We manually review 12 biography pages containing a total of 103 interaction descriptions. Our extraction pipeline could capture at least 94.00% of the interactions across these pages. Unidentified interactions are often due to certain interactions require multiple segments of text to infer, making their recognition hard even for humans.

4.2.2 Annotating

After we extract candidate quadruples using the above method, we randomly select 4,507 quadruples for manual annotation. Furthermore, we also split each interaction quadruple (*Person1*, *Person2*, *Time*, *Location*) into two trajectory triples (*Person*, *Time*, *Location*) for the annotation of trajectory category, since the detection of trajectory is an auxiliary task in our multi-task learning framework. Our annotation has involved three undergraduate students: two annotators and one checker. We annotate 4,507 interaction quadruples and 9,014 trajectory triplets. The distribution of positive and negative samples is shown in Table 1, with positive indicating actual interactions or trajectories, and negative indicating otherwise.

Type	Positive	Negative	Total
Interaction	2,351	2,156	4,507
Trajectory	5,730	3,284	9,014

Table 1: Distribution of the $WikiInteraction$ Dataset.

5 Method

Our FALCON framework (Fig. 2) processes candidate interactions comprising a text segment s and a quadruple q. First, q is decomposed into two trajectory triplets (t_1, t_2). The quadruple q drives the main **Interaction Classification** task, while the triplets t_1 and t_2 drive the auxiliary **Trajectory Classification** task.

We employ a trainable AR-BERT model as the primary feature extractor for both tasks, yielding interaction features H'_{inter} and trajectory features H'_{tra}. Simultaneously, a *frozen* AR-BERT model, pre-trained solely on trajectory data, extracts trajectory features H_{tra}. Transfer learning is incorporated by fusing H'_{inter} and H_{tra} into H'_{fusion} for the final interaction prediction. The model is jointly trained using multi-task learning.

5.1 AR-BERT

The architecture of AR-BERT, designed to enhance BERT's understanding of entities in a given text.

Input and Special Markers. Given an input text segment s and a set of entities $E = \{e_1, e_2, \ldots, e_n\}$, we insert special tokens to enhance entity capture. Specifically, we use '#' for *Person1*, '$' for *Person2*, '*' for *Time*, and '&' for *Location*. The modified sentence is fed into BERT, where the final hidden state of the `[CLS]` token serves as the global representation.

Entity Information Representation. We employ a multi-step process to represent entities. First, we obtain the vector for the k-th occurrence of entity e_i via mean pooling:

$$H_i^k = \frac{1}{d-c+1} \sum_{t=c}^{d} \mathbf{H}_t, \tag{1}$$

where c and d denote the start and end positions of the entity. Next, an attention mechanism fuses these occurrences. The importance weight w_i^k and normalized score δ_i^k are computed as:

$$w_i^k = \tanh\left(\boldsymbol{W}_{attn} H_i^k + \boldsymbol{b}_{attn}\right), \tag{2}$$

$$\delta_i^k = \frac{\exp(w_i^k)}{\sum_{u=1}^{v} \exp(w_i^u)}. \tag{3}$$

The final entity embedding is the weighted sum:

$$H_i = \sum_{k=1}^{v} \delta_i^k H_i^k. \tag{4}$$

Feature Concatenation. Each entity embedding H_{e_i} and the global sentence representation $\mathbf{H}_0$ (the hidden state of the `[CLS]` token) pass through a fully connected layer with activation:

$$H'_{e_i} = \mathbf{W}_i \left(\tanh(H_{e_i})\right) + \mathbf{b}_i, \tag{5}$$

$$H'_0 = \mathbf{W}_0 \left(\tanh(\mathbf{H}_0)\right) + \mathbf{b}_0. \tag{6}$$

Finally, we concatenate them to form the output feature:

$$\begin{aligned} H' &= \mathbf{AR_BERT}(s, E) \\ &= \text{concat}\left(H'_0, H'_{e_1}, \ldots, H'_{e_n}\right), \end{aligned} \tag{7}$$

where $H' \in R^{(n+1)d}$.

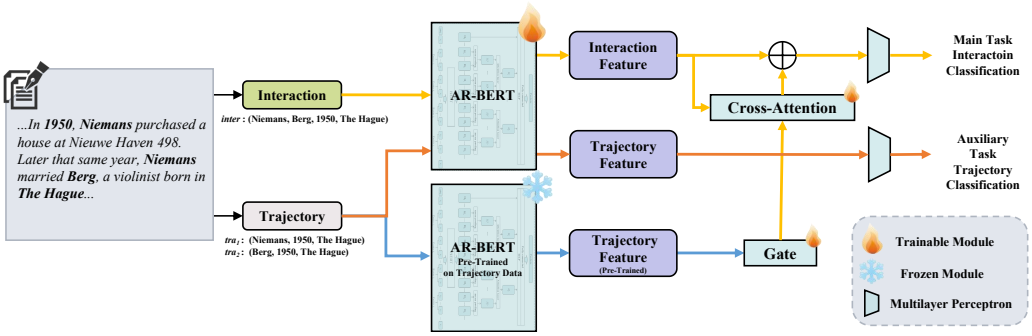


Figure 2: The framework of our FALCON method. The candidate interaction is decomposed into trajectory triplets. The main task handles interaction classification, while the auxiliary task handles trajectory classification via multi-task learning.

5.2 Main Task: Interaction Classification

The input consists of text s_{inter} and quadruple E_q. The base interaction feature is extracted as $H'_{inter} = \mathbf{AR_BERT}(s_{inter}, E_q) \in R^{5d}$.

Feature Transfer. To leverage trajectory knowledge, we utilize a frozen trajectory extractor f_{tra}, which is built upon the AR-BERT architecture yet pre-trained on a separate trajectory dataset (Zhang et al., 2025). Crucially, we decompose the interaction quadruple $E_q = (Person_1, Person_2, Time, Location)$ into two constituent trajectory triplets: $E_{t1} = (Person_1, Time, Location)$ and $E_{t2} = (Person_2, Time, Location)$. These triplets are individually fed into the extractor to obtain independent trajectory features $H^t_{\text{tra}_i} = f_{\text{tra}}(s_{inter}, E_{ti}) \in R^d$ for $i \in \{1, 2\}$, which capture the spatio-temporal validity of each participant.

We integrate these features using a **Gating Mechanism** followed by **Cross-Attention**. The gating filters the trajectory information:

$$\mathbf{gate}_{\text{tra}_i} = \sigma\left(\mathbf{W}_{\text{gate}} H^t_{\text{tra}_i}\right), \tag{8}$$

$$H^g_{\text{tra}_i} = \mathbf{gate}_{\text{tra}_i} \odot H^t_{\text{tra}_i}, \tag{9}$$

where $\mathbf{W}_{\text{gate}} \in R^{d \times d}$. Subsequently, cross-attention is applied using H'_{inter} as the query source:

$$Q = \mathbf{W}_Q H'_{\text{inter}}, \tag{10}$$

$$H^a_{\text{tra}_i} = \sigma\left(\frac{Q(H^g_{\text{tra}_i})^\top}{\sqrt{d}}\right) H^g_{\text{tra}_i}, \tag{11}$$

where $\mathbf{W}_Q \in R^{d \times 5d}$. The final fused feature H'_{fusion} is obtained by concatenating the interaction feature with the attended trajectory features:

$$H'_{fusion} = \text{concat}\left(H'_{\text{inter}}, H^a_{\text{tra}_1}, H^a_{\text{tra}_2}\right). \tag{12}$$

Loss Function. The prediction is $\hat{y}_{inter} = \text{Softmax}(W_{inter} H'_{\text{fusion}})$. We use the cross-entropy loss:

$$\mathcal{L}_{\text{inter}} = -\sum_{j=1}^{J} \Big[y^j_{inter} \log(\hat{y}^j_{inter}) + (1 - y^j_{inter}) \log(1 - \hat{y}^j_{inter}) \Big]. \tag{13}$$

5.3 Auxiliary Task: Trajectory Classification

The interaction data provides two auxiliary trajectory samples per instance. Features are extracted via AR-BERT as $H'_{tra_i} = \mathbf{AR_BERT}(t_{tra_i}, E_{t_i})$. The prediction is $\hat{y}_{tra_i} = \text{Softmax}(W_{tra} H'_{\text{tra}_i})$. The auxiliary loss is the average of the two trajectory losses:

$$\mathcal{L}_{\text{tra}_i} = -\sum_{j=1}^{J} \Big[y^j_{tra_i} \log(\hat{y}^j_{tra_i}) + (1 - y^j_{tra_i}) \log(1 - \hat{y}^j_{tra_i}) \Big], \tag{14}$$

$$\mathcal{L}_{\text{tra}} = \frac{1}{2}(\mathcal{L}_{\text{tra}_1} + \mathcal{L}_{\text{tra}_2}). \tag{15}$$

5.4 Multi-Task Learning

We jointly train the model using adaptive weighting (Liebel and Körner, 2018) to balance the tasks:

$$\mathcal{L} = \frac{1}{2c_1^2}\mathcal{L}_{\text{inter}} + \frac{1}{2c_2^2}\mathcal{L}_{\text{tra}} + \log\left(1 + c_1^2\right) + \log\left(1 + c_2^2\right), \tag{16}$$

where c_1 and c_2 are learnable parameters initialized to 1.

6 Experiments

6.1 Train/Test Split

For data partitioning, we adhere to the principle of isolating by "biography pages" and divide the

Type	Methods	Wiki (In-Domain)				Britannica (Out-of-Domain)				News (Out-of-Domain)			
		Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
LLMs	GPT-4o-mini (zero-shot)	74.17	72.69	80.60	76.44	73.85	72.10	80.15	75.91	73.12	71.88	79.40	75.45
	GPT-4o-mini (few-shot)	79.15	76.12	85.47	80.53	75.88	71.93	82.97	77.06	75.05	71.40	82.38	76.50
	DeepSeek-V3 (zero-shot)	78.85	74.23	85.98	79.67	77.40	72.48	84.46	78.01	76.90	72.34	83.47	77.51
	DeepSeek-V3 (few-shot)	82.05	78.18	89.17	83.31	79.10	73.92	86.54	79.73	79.20	74.50	85.45	79.60
	GPT-o1 (zero-shot)	79.95	77.09	84.53	80.64	79.82	76.55	84.10	80.15	79.45	76.12	83.85	79.81
	GPT-o1 (few-shot)	83.50	79.13	**91.76**	84.98	81.05	75.68	**87.96**	81.37	80.60	75.60	85.72	80.34
	Claude 4 Sonnet (zero-shot)	80.42	77.15	86.35	81.49	79.15	75.30	84.20	79.50	78.80	75.10	83.90	79.26
	Claude 4 Sonnet (few-shot)	83.95	80.12	90.78	85.12	80.50	76.20	87.05	81.26	79.85	75.56	**85.95**	80.42
SLMs	BERT	76.61	73.33	84.43	78.49	73.02	70.88	79.36	74.88	72.45	70.21	78.64	74.18
	RoBERTa	80.16	74.87	89.55	81.55	75.18	69.75	83.42	75.97	74.62	69.12	82.51	75.22
	COSMOS	81.93	81.01	85.73	83.30	78.21	78.02	79.14	78.58	77.61	77.10	78.55	77.82
	R-Bert	82.37	79.69	88.70	84.01	78.47	76.95	82.01	79.40	77.98	76.30	81.69	78.90
	FALCON (ours)	**85.48**	**83.67**	89.55	**86.51**	**82.54**	**82.33**	83.19	**82.76**	**80.78**	**80.58**	81.67	**81.12**

Table 2: Models are trained on Wiki and evaluated on Wiki, Britannica, and News (Acc, P, R, and F1 in %). The methods are categorized into Large Language Models (LLMs) and Small Language Models (SLMs).

primary *WikiInteraction* dataset into training, validation, and test sets in a 7:1:2 ratio. This approach ensures that information pertaining to any given individual appears exclusively within a single data subset, thereby strictly preventing data leakage. **To further evaluate the model's cross-domain applicability, we additionally annotated 500 samples from Encyclopedia Britannica[2] and 500 from news archives[3].**

6.2 Evaluation Metrics

To quantitatively evaluate our model, we assess its classification performance on the interaction task by computing Accuracy (Acc), Precision (P), Recall (R), and F1-score (F1).

6.3 Baseline Methods

To comprehensively evaluate the effectiveness of FALCON, we compare it against a diverse set of baselines, categorized into Small Language Models (SLMs) and Large Language Models (LLMs). For SLMs, we employ **BERT** (Devlin, 2018) as a standard pretraining baseline and **RoBERTa** (Liu, 2019), which optimizes representation learning via dynamic masking and removing the NSP objective. We also include **R-Bert** (Wu and He, 2019), which enhances entity context by explicitly marking target entities and integrating their position-aware representations. Additionally, we adapt **COSMOS** (Zhang et al., 2025), a trajectory-focused model that does not explicitly model entity relationships, by retraining it specifically for the interaction quadruple extraction task. Regarding LLMs, we assess a broader range of representative models: the lightweight **GPT-4o-mini**[4] to evaluate generative reasoning, the more advanced **GPT-o1**[5], as well as **Claude 4 Sonnet**[6] and **DeepSeek-V3**[7]. We evaluate these models under both zero-shot (utilizing Chain-of-Thought) and Few-shot settings. For the Few-shot setting, we employ a Retrieval-Augmented Generation (RAG) approach to retrieve five relevant demonstrations from the *WikiInteraction* training set (prompt details in Appendix A.3).

6.4 Experimental Results

Table 2 presents the evaluation results across three datasets. FALCON consistently achieves the highest F1 scores: 86.51% on Wiki, 82.76% on Britannica, and 81.12% on News, outperforming all baselines across both in-domain and out-of-domain settings. Overall, the large language model Claude 4 Sonnet (few-shot) ranks second, while the small model R-BERT ranks third among all evaluated approaches.

Comparison with LLMs. FALCON demonstrates consistent improvements over state-of-the-art LLMs, with gains of 1.39-1.53% on Wiki and 0.70-1.39% on out-of-domain datasets. The advantage primarily comes from higher precision (83.67% vs. 79-80%), with FALCON achieving 3-4% better precision while trading off only 1-2% in recall compared to LLMs. Beyond performance, using LLMs for large-scale extraction is

[2]https://www.britannica.com/
[3]https://huggingface.co/datasets/rjac/all-the-news-2-1-Component-one
[4]https://platform.openai.com/docs/models#gpt-4o-mini
[5]https://platform.openai.com/docs/models/o1
[6]https://www.anthropic.com/news/claude-4
[7]https://github.com/deepseek-ai/DeepSeek-V3

prohibitively expensive. For instance, processing all Wikipedia biographies with GPT-o1 costs over $1.2 million, while our candidate classification approach still incurs approximately $157k (details in Appendix A.5). In contrast, FALCON achieves this with negligible GPU costs.

Comparison with small models. FALCON shows larger improvements over small models, outperforming R-BERT by 2.50% on Wiki, 3.36% on Britannica, and 2.22% on News. Despite sharing the same BERT base, this significant margin indicates the contribution of the proposed components.

6.5 Ablation Study

We conduct an ablation study on the *WikiInteraction* benchmark to evaluate the effectiveness of key components, as shown in Table 3. Specifically, FALCON$_{w/o\ ft}$ excludes the feature transfer module, FALCON$_{w/o\ mt}$ omits the multi-task learning strategy, and FALCON$_{w/o\ ft\&mt}$ removes both components. Consistent with our expectations, FALCON achieves the best performance, while FALCON$_{w/o\ ft\&mt}$ exhibits the worst performance, underscoring the effectiveness of each component. Notably, FALCON$_{w/o\ mt}$ achieves the highest Recall. This suggests that the multi-task learning mechanism enforces stricter spatio-temporal verification; while this causes a slight reduction in Recall, it yields a substantial gain in Precision, resulting in superior overall performance.

Additionally, we conduct detailed ablations on the design of multi-task learning and feature transfer in the Appendix A.2.

Methods	Acc (%)	P (%)	R (%)	F1 (%)
$FALCON_{w/o\ ft\&mt}$	83.70	82.33	87.42	84.80
$FALCON_{w/o\ mt}$	83.81	80.76	**90.40**	85.31
$FALCON_{w/o\ ft}$	84.70	82.26	89.98	85.94
$FALCON$	**85.48**	**83.67**	89.55	**86.51**

Table 3: Results of the ablation study.

7 Analysis of Extracted Interactions

We extract **658,966** spatio-temporal interaction quadruples from **4,226,049** candidates generated from English Wikipedia. Manual evaluation of 300 extracted samples showed an accuracy of 82% (246 quadruples). Conversely, inspection of 100 samples labeled as negative by FALCON confirmed that 93% were indeed incorrect, aligning with the model's test set performance.

To ensure data uniqueness and spatial precision for downstream analysis, we implemented a rigorous post-processing pipeline involving URL-based entity deduplication and coordinate standardization (detailed protocols and algorithms are provided in Appendix A.4).

Consequently, we obtained a refined dataset of 293,586 distinct quadruples where each individual possesses a separate Wikipedia page. This subset spans from 1000 to 2024[8], comprising 291,136 unique individuals and 49,579 locations.

In the Appendix A.8, we present the complete dataset from the perspectives of geography and spatio-temporal networks. In the following section, we demonstrate the data's utility by analyzing interactions among US political figures, offering a new perspective on political polarization. Additionally, we extracted 77,836 raw interaction quadruples from news data (see Appendix A.9 for details).

7.1 Case Study: Long-term Analysis of US Political Polarization

7.1.1 Data Source and Classification

We focus on the real-world interactions of US political figures extracted from Wikipedia to examine political polarization. Utilizing our extracted data, we identify 14,084 interactions among 3,896 Republicans and 3,995 Democrats (1960–2024). These interactions are categorized as intra-party, where individuals belong to the same party, or inter-party, where they come from different parties.

We classify interactions into three types using GPT-4o-mini, achieving 93.50% accuracy on a manually verified subset of 200 samples (prompt details in Appendix A.6): *Adversarial* (conflicting interests, e.g., debates), *Cooperative* (joint actions, e.g., legislation), and *Neutral* (symbolic events, e.g., ceremonies). While *Adversarial* and *Cooperative* align with established political frameworks (Bassan-Nygate and Weiss, 2022; Jost et al., 2022; Bendix and MacKay, 2017), *Neutral* encompasses additional interactions beyond these categories, as illustrated above. Table 4 details the final distribution.

Type	Intra-party	Inter-party
Cooperative	5,182	1,670
Adversarial	1,634	2,317
Neutral	1,468	1,812

Table 4: Detailed statistics of interaction types classified by GPT-4o-mini.

[8]Wikipedia dump: January 11, 2025 (https://dumps.wikimedia.org/).

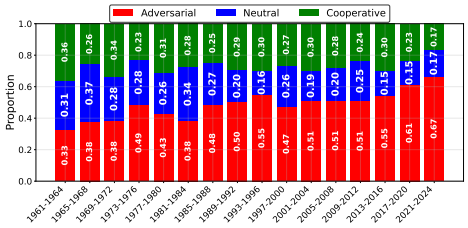


Figure 3: Evolution of inter-party interaction types.

7.1.2 Rising Adversarial Interactions

In terms of inter-party dynamics, we observe that the proportion of inter-party interactions has dropped significantly from 47% (1960s) to 24% (2020s). Notably, within this reduced volume, *Adversarial* interactions have surged from 32.8% to 66.7% (Fig.3), while *Cooperative* and *Neutral* types have declined. This trend aligns with observations of growing polarization in the US (Fiorina and Abrams, 2008).

7.1.3 Network-Based Polarization and Regional Divergence

To further capture the complexity of political interactions, we examine polarization by constructing interaction networks, referencing the classic study on political polarization (Conover et al., 2011).

In our constructed network, each node represents a political figure, and edges denote interactions. Adopting the *Weighted Signed Networks* (WSN) (Kumar et al., 2016), we assign signed weights to edges to reflect the nature of the interactions (refer to the Appendix A.7 for specific weight assignments). Figure 4 visualizes the cumulative network spanning 1960–2024 using the ForceAtlas2 algorithm. In this visualization, nodes are color-coded by political party, with red representing Republicans and blue representing Democrats, clearly revealing distinct partisan clusters.

To quantify polarization over time, we compute *standardized modularity* (details in Appendix A.7). Longitudinal results (Figure 5a) confirm a continuous rise in national polarization, with the sharpest acceleration occurring during the Obama and Trump administrations. These findings align with poll-based conclusions that partisan divisions have reached record levels during these periods (Doherty et al., 2017). However, regional trends vary significantly (Figure 5b). Massachusetts shows consistent growth to lead in 2024, while California presents a unique case with recent declines. Most strikingly, Washington D.C. demonstrates the steepest increase among major political centers, rising from the lowest initial polarization to the second highest.

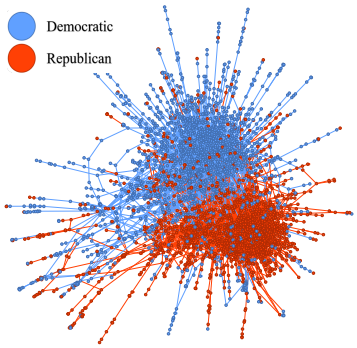


Figure 4: US political interaction network (1960-2024). Edges are weighted: Neutral (+1), Cooperative (+2), Adversarial (-2).

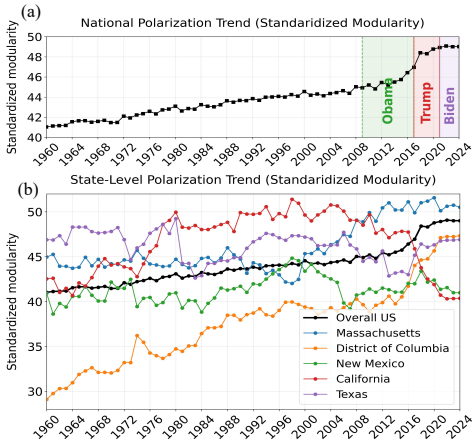


Figure 5: (a) National polarization trend; (b) State-level trends.

8 Conclusion

We propose a new task of extracting spatio-temporal interactions from unstructured text and introduce FALCON. By integrating AR-Bert, feature transfer, and multi-task learning, FALCON provides an efficient and effective solution for large-scale extraction of human spatio-temporal interactions. Although trained on Wikipedia, FALCON demonstrates broad cross-domain applicability, extending effectively to Encyclopedia Britannica and News datasets. To validate the method and showcase the potential of the extracted data, we analyze US political interactions, offering a new perspective on political polarization. We hope that the open-sourced code, the extracted interactions, and the *WikiInteraction* ground truth dataset, can support the spatio-temporal interaction extraction research and the analytical studies based on these spatio-temporal interactions. As the largest of its kind, our dataset can be the basis for data-driven grand narratives and explorations of human interaction mechanisms.

Limitations

Despite the effectiveness of FALCON in extracting spatio-temporal interactions, several limitations remain:

- **Dependency on External LLMs:** While FALCON extracts interactions efficiently, the subsequent classification of interaction types relies on GPT-4o-mini. This introduces a dependency on external APIs, which future work aims to address with a dedicated end-to-end architecture.
- **Source Bias:** Our reliance on English Wikipedia inherits its inherent Western-centric biases. Consequently, the extracted networks potentially underrepresent figures from non-English speaking regions or marginalized communities.
- **Language Specificity:** The model is trained and evaluated solely on English corpora. Although the architecture is language-agnostic, its generalization to other languages remains untested.

Ethics Statement

Our study uses the English Wikipedia and public news archives, and we ensure that our data collection process does not violate any privacy or confidentiality concerns. A potential ethical concern is the misuse to extract spatio-temporal interactions of individual users (non-famous people). However, since our framework relies on detailed textual descriptions of public events, the risk arises from the availability of personal information in such descriptions, rather than the framework itself. Therefore, we believe that there are no essential ethical questions raised by our study.

References

Apoorv Agarwal. 2016. Social Network Extraction from Text. Ph.D. thesis, Columbia University.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. Journal of economic perspectives, 31(2):211–236.

Lars Backstrom, Daniel P. Huttenlocher, Jon M. Kleinberg, and Xiangyang Lan. 2006. Group formation in large social networks: membership, growth, and evolution. In Knowledge Discovery and Data Mining.

Albert-Laszlo Barabâsi, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. 2002. Evolution of the social network of scientific collaborations. Physica A: Statistical mechanics and its applications, 311(3-4):590–614.

Lotem Bassan-Nygate and Chagai M Weiss. 2022. Party competition and cooperation shape affective polarization: evidence from natural and survey experiments in israel. Comparative Political Studies, 55(2):287–318.

William Bendix and Jon MacKay. 2017. Partisan infighting among house republicans: leaders, factions, and networks of interests. Legislative Studies Quarterly, 42(4):549–577.

Katy Börner, Jeegar T Maru, and Robert L Goldstone. 2004. The simultaneous evolution of author and paper networks. Proceedings of the National Academy of Sciences, 101(suppl_1):5266–5273.

Damon Centola. 2010. The spread of behavior in an online social network experiment. science, 329(5996):1194–1197.

Lihu Chen and Gaël Varoquaux. 2025. What is the role of small models in the llm era: A survey. Preprint, arXiv:2409.06857.

Federico Cinus, Marco Minici, Luca Luceri, and Emilio Ferrara. 2025. Exposing cross-platform coordinated inauthentic activity in the run-up to the 2024 us election. In Proceedings of the ACM on Web Conference 2025, pages 541–559.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In Proceedings of the international aaai conference on web and social media, volume 5, pages 89–96.

David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. 2010. Inferring social ties from geographic coincidences. Proceedings of the National Academy of Sciences, 107(52):22436–22441.

Cesi Cruz, Julien Labonne, and Pablo Querubin. 2017. Politician family networks and electoral outcomes: Evidence from the philippines. American Economic Review, 107(10):3006–3037.

Ashok Deb, Luca Luceri, Adam Badaway, and Emilio Ferrara. 2019. Perils and challenges of social media and election manipulation analysis: The 2018 us midterms. In Companion proceedings of the 2019 world wide web conference, pages 237–247.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Carroll Doherty, Jocelyn Kiley, and Bridget Johnson. 2017. The partisan divide on political values grows even wider. Pew Research Center, 5.

Tim Donkers and Jürgen Ziegler. 2025. Understanding online polarization through human-agent interaction in a synthetic llm-based social network. In Proceedings of the International AAAI Conference on Web and Social Media, volume 19, pages 457–478.

Morris P Fiorina and Samuel J Abrams. 2008. Political polarization in the american public. Annu. Rev. Polit. Sci., 11(1):563–588.

James Flamino, Alessandro Galeazzi, Stuart Feldman, Michael W Macy, Brendan Cross, Zhenkun Zhou, Matteo Serafino, Alexandre Bovet, Hernán A Makse, and Boleslaw K Szymanski. 2023. Political polarization of news media and influencers on twitter in the 2016 and 2020 us presidential elections. Nature Human Behaviour, 7(6):904–916.

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, and 1 others. 2018. Science of science. Science, 359(6379):eaao0185.

Vyacheslav Fos, Elisabeth Kempf, and Margarita Tsoutsoura. 2022. The political polarization of corporate america. Technical report, National Bureau of Economic Research.

Michael A Fuller and Hongsu Wang. 2021. Structuring, recording, and analyzing historical networks in the china biographical database. Journal of Historical Network Research, 5(1):248–270.

Venkata Rama Kiran Garimella and Ingmar Weber. 2017. A long-term analysis of polarization on twitter. In Proceedings of the International AAAI Conference on Web and social media, volume 11, pages 528–531.

Olivier Gergaud, Morgane Laouenan, and Etienne Wasmer. 2016. A brief history of human time. exploring a database of "notable people"'. LIEPP Working Paper, (46).

Chien-leng Hsu and Han Woo Park. 2012. Mapping online social networks of korean politicians. Government information quarterly, 29(2):169–181.

Johannes Illenberger, Kai Nagel, and Gunnar Flötteröd. 2013. The role of spatial interaction in social networks. Networks and Spatial Economics, 13:255–282.

Matthew O Jackson. 2011. An overview of social networks and economic applications. Handbook of social economics, 1:511–585.

John T Jost, Delia S Baldassarri, and James N Druckman. 2022. Cognitive–motivational mechanisms of political polarization in social-communicative contexts. Nature reviews psychology, 1(10):560–576.

Jon Kleinberg. 2013. Analysis of large-scale social and information networks. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1987):20120378.

Emily Kubin and Christian Von Sikorski. 2021. The role of (social) media in political polarization: a systematic review. Annals of the International Communication Association, 45(3):188–206.

Srijan Kumar, Francesca Spezzano, VS Subrahmanian, and Christos Faloutsos. 2016. Edge weight prediction in weighted signed networks. In 2016 IEEE 16th international conference on data mining (ICDM), pages 221–230. IEEE.

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. ACM Computing Surveys (CSUR), 52(5):1–40.

Lukas Liebel and Marco Körner. 2018. Auxiliary tasks in multi-task learning. arXiv preprint arXiv:1805.06334.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 364.

Christoph Lutz and Christian Pieter Hoffmann. 2017. The dark side of online participation: exploring non-, passive and negative participation. Information, Communication & Society, 20(6):876–897.

Cedric Möller, Jens Lehmann, and Ricardo Usbeck. 2022. Survey on english entity linking on wikidata: Datasets and approaches. Semantic Web, 13(6):925–966.

Mark EJ Newman. 2001. Scientific collaboration networks. i. network construction and fundamental results. Physical review E, 64(1):016131.

M Nijila and M T Kala. 2018. Extraction of relationship between characters in narrative summaries. In 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR), pages 1–5.

Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. 2021. Coordinated behavior on social media in 2019 uk general election. In Proceedings of the international AAAI conference on web and social media, volume 15, pages 443–454.

Motasem S Obeidat, Md Sultan Al Nahian, and Ramakanth Kavuluru. 2025. Do llms surpass encoders for biomedical ner? In 2025 IEEE 13th International Conference on Healthcare Informatics (ICHI), pages 352–358.

Lindsay O'Neill. 2014. The opened letter: Networking in the Early modern British world. University of Pennsylvania Press.

Jukka-Pekka Onnela, Samuel Arbesman, Marta C González, Albert-László Barabási, and Nicholas A Christakis. 2011. Geographic constraints on social network groups. PLoS one, 6(4):e16939.

John F Padgett and Christopher K Ansell. 1993. Robust action and the rise of the medici, 1400-1434. American journal of sociology, 98(6):1259–1319.

Pew Research Center. 2017. The partisan divide on political values grows even wider. Trust, facts and democracy.

Pew Research Center. 2019. Partisan antipathy: More intense, more personal. Pew Research Center.

Thomas Plotkowiak and Katarina Stanoevska-Slabeva. 2013. German politicians and their twitter networks in the bundestag election 2009. First Monday.

Neeti Pokhriyal, Benjamin A Valentino, and Soroush Vosoughi. 2023. Quantifying participation biases on social media. EPJ Data Science, 12(1):26.

Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. 2011. Influence and passivity in social media. In Proceedings of the 20th international conference companion on World wide web, pages 113–114.

Yizhou Sun, Rick Barber, Manish Gupta, Charu C Aggarwal, and Jiawei Han. 2011. Co-author relationship prediction in heterogeneous bibliographic networks. In 2011 International Conference on Advances in Social Networks Analysis and Mining, pages 121–128. IEEE.

Jie Tang, Duo Zhang, and Limin Yao. 2007. Social network extraction of academic researchers. In Seventh IEEE International Conference on Data Mining (ICDM 2007), pages 292–301. IEEE.

Lu Tao and Shunxiang Zhang. 2020. Character relationship extraction method based on bilstm. In International Conference on Applications and Techniques in Cyber Intelligence ATCI 2019: Applications and Techniques in Cyber Intelligence 7, pages 1432–1441. Springer.

Zeynep Tufekci. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In Proceedings of the international AAAI conference on web and social media, volume 8, pages 505–514.

Brian Uzzi and Jarrett Spiro. 2005. Collaboration and creativity: The small world problem. American journal of sociology, 111(2):447–504.

Sander Van der Linden, Costas Panagopoulos, and Jon Roozenbeek. 2020. You are fake news: political bias in perceptions of fake news. Media, culture & society, 42(3):460–470.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In Proceedings of the 28th ACM international conference on information and knowledge management, pages 2361–2364.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. Frontiers of Computer Science, 18(6):186357.

Ying Zhang, Xiaofeng Li, Zhaoyang Liu, and Haipeng Zhang. 2024. Paths of a million people: Extracting life trajectories from wikipedia. arXiv preprint arXiv:2406.00032.

Ying Zhang, Xiaofeng Li, Zhaoyang Liu, and Haipeng Zhang. 2025. Paths of a million people: Extracting life trajectories from wikipedia. In Proceedings of the International AAAI Conference on Web and Social Media, volume 19, pages 2226–2240.

A Appendix

A.1 Implementation Details

In this paper, we use a BERT-base[9] model to generate word embeddings. While any suitable model can be employed as well, in FALCON , we set $d = 768$. We train FALCON using the AdamW optimizer with a learning rate configuration of $5e^{-5}$. All the aforementioned experiments are conducted on two RTX 3090 GPUs.

A.2 Detailed Ablation

Methods	Acc (%)	P (%)	R (%)	F1 (%)
$FALCON_{w/o\ aw}$	84.26	81.20	89.34	85.51
$FALCON_{concat}$	85.25	**84.85**	87.21	86.01
$FALCON$	**85.48**	83.67	**89.55**	**86.51**

Table 5: Specific ablation study.

We conduct specific ablation studies on multi-task learning and feature transfer, as shown in Table 5. For multi-task learning, we validate the effectiveness of the adaptive task weighting strategy. Specifically, $\text{FALCON}_{\text{w/o aw}}$ denotes the model without the adaptive task weighting strategy. The results show a significant decrease in overall metrics. The adaptive weighting strategy enables the model to automatically focus on more challenging interaction tasks, thereby enhancing its performance on this task. Regarding feature transfer, we assess the effectiveness of the feature fusion module. $\text{FALCON}_{\text{concat}}$ represents the model where the feature fusion module is removed and trajectory features and interaction features are merely concatenated. We observe that compared to $\text{FALCON}_{\text{w/o ft}}$ (refer to Table 3), which does not utilize any feature transfer strategy, $\text{FALCON}_{\text{concat}}$ only achieves a marginal improvement in the F1 score. Although precision increases by 2.59%, recall decreases by 2.77%. This indicates that the concatenation strategy in our task trades recall for precision. In contrast, FALCON, which employs a feature fusion module in its feature transfer strategy, manages to robustly increase precision while only slightly sacrificing recall, thereby achieving better improvements in the overall metrics of F1 and recall.

[9]https://github.com/google-research/bert

A.3 Prompt for LLM Baseline

Figure 6 shows the prompt provided when we use LLMs. The temperature here is set to 0 and one trial is performed for each input.

Prompt:

You will be provided with a text and a quadruple of entities (*Person1*, *Person2*, *Time*, *Location*). If *Person1* and *Person2* are both present at the specified *Time* and *Location*, and there is an association between *Person1* and *Person2*, we consider that an interaction has occurred between them.

Your task is to answer ``Yes" or ``No" to indicate whether the interaction occurred. No additional explanations or responses are required.

Text Content: {*Text*}

Quadruple: {*Person1*, *Person2*, *Time*, *Location*}

Reasoning Steps:

1. Identify the presence of *Person1* at the specified *Time* and *Location*.
2. Identify the presence of *Person2* at the specified *Time* and *Location*.
3. Determine if there is a association between *Person1* and *Person2*.
4. If all the above conditions are satisfied, answer "Yes". Otherwise, answer "No".

Is the interaction present? Yes or No.

Figure 6: The zero-shot prompt for LLM Baseline.

Prompt:

You will be provided with a text and a quadruple of entities (*Person1, Person2, Time, Location*). If *Person1* and *Person2* are both present at the specified *Time* and *Location*, and there is an association between *Person1* and *Person2*, we consider that an interaction has occurred between them.

Your task is to answer "Yes" or "No" to indicate whether the interaction occurred. No additional explanations or responses are required.

Reasoning Steps:

1. Identify the presence of *Person1* at the specified *Time* and *Location*.
2. Identify the presence of *Person2* at the specified *Time* and *Location*.
3. Determine if there is an association between *Person1* and *Person2*.
4. If all the above conditions are satisfied, answer "Yes". Otherwise, answer "No".

Examples: {*Examples*}

Now, analyze the following new case:

Text Content: {*Text*}

Quadruple: {*Person1,Person2,Time,Location*}

Is the interaction present? Yes or No.

Figure 7: The few-shot prompt for LLM Baseline.

A.4 Data Post-Processing Protocols

To ensure the reliability of the extracted interactions, we implemented the following protocols:

- **Entity Deduplication:** To address naming ambiguities and duplicate records arising from mutual references (i.e., the same interaction appearing on both participants' pages), we utilized the unique Wikipedia URL of each entity as a distinct identifier. This enabled rigorous deduplication by merging entries sharing identical entity IDs, timestamps, and locations.

- **Spatial Standardization:** To mitigate location misclassification and ensure spatial resolution, we employed a geocoding service[10] to standardize location names into precise coordinates.

[10]https://developers.google.com/maps/documentation/geocoding

A.5 Cost Analysis of Large Language Models

Taking GPT-o1 as an example, we provide a detailed cost analysis. Our calculation involves two components: deterministic input costs based on dataset size, and estimated inference costs (reasoning + generation) derived from pilot experiments.

A.5.1 Methodology

We calculate the **Input Cost** using the standard pricing of $15.00 per 1 million input tokens. However, since the costs for reasoning and output tokens are dynamic and priced significantly higher (e.g., $60.00 per 1M tokens for GPT-o1), we estimate them through a sampling strategy:

1. We randomly sample 20 cases for each scenario and process them via the API to obtain the *Total Cost* per case.
2. For each case, we calculate the *Input Cost* based on its input token count.
3. The **Inference Cost** (covering both reasoning and output tokens) is derived by subtracting the calculated Input Cost from the Total Cost:

$$\text{Cost}_{\text{inf}} = \text{Cost}_{\text{total}} - \text{Cost}_{\text{in}} \tag{17}$$

4. We then calculate the average Inference Cost per case and project it to the full dataset.

A.5.2 Full Corpus Extraction

This scenario corresponds to the direct processing of the entire English Wikipedia biography corpus.

- **Dataset Scope**: The corpus consists of 1,930,519 biography pages.
- **Input Cost (Calculated)**: Processing the raw text involves a total of 71,290,205,632 tokens.

$$\begin{aligned}\text{Cost}_{\text{in}} &= \frac{71,290,205,632}{1,000,000} \times \$15.00 \\ &\approx \$1,069,353.08\end{aligned} \tag{18}$$

- **Total Cost Projection**: Based on our pilot study of 20 documents:
  - *Avg. Inference Cost*: After subtracting input costs, the residual cost for reasoning and generation averages **$0.102 per document**.
  - *Projected Inference Cost*:

$$\text{Cost}_{\text{inf}} = 1,930,519 \times \$0.102 \approx \$196,912.94 \tag{19}$$

- **Estimated Total**: The aggregate cost rises to approximately **$1,266,266.02**. In this scenario, due to the massive length of raw texts, input processing remains the dominant cost factor.

A.5.3 Candidate Quadruple Classification

This scenario corresponds to our experimental baseline setting, where the model classifies pre-extracted candidate quadruples using the Few-Shot strategy.

- **Workload**: The heuristic method generates 4,226,049 candidate interaction quadruplets.
- **Input Cost (Calculated)**: Including the system prompt, context, and few-shot demonstrations, the total input volume is 5,714,640,772 tokens.

$$\begin{aligned}\text{Cost}_{\text{in}} &= \frac{5,714,640,772}{1,000,000} \times \$15.00 \\ &\approx \$85,719.61\end{aligned} \tag{20}$$

- **Total Cost Projection**: Based on the pilot study of 20 candidate samples:
  - *Avg. Inference Cost*: Using the subtraction method ($\text{Cost}_{\text{total}} - \text{Cost}_{\text{in}}$), the average cost for reasoning and verification logic is **$0.017 per candidate**.
  - *Projected Inference Cost*:

$$\text{Cost}_{\text{inf}} = 4,226,049 \times \$0.017 \approx \$71,842.83 \tag{21}$$

- **Estimated Total**: The aggregate cost is approximately **$157,562.44**. Although our proposed problem formulation—transforming the extraction task into a candidate classification problem—significantly reduces the token consumption compared to full corpus processing, the expense remains prohibitively high for LLMs.

A.6 Interaction Classification Prompt

Figure 8 presents the specific prompt used by GPT-4o-mini to categorize political interactions into *Adversarial*, *Cooperative*, and *Neutral* types, as discussed in the case study.

Prompt:
You will be provided with a text, a quadruple of *entities (Person1, Person2, Time, Location)* and the political party information of *Person1* and *Person2*.
Please analyze the interaction between *Person 1* and *Person 2* :

1. Understand this interaction
2. Categorize this interaction with definitions:
   - Cooperative: Joint actions towards common political goals (e.g., political work cooperation, face-to-face supportt)
   - Adversarial: Conflicting political interests (e.g., election competition, debate)
   - Neutral: Non-political/symbolic interactions (e.g., personal relationships, ceremonial meetings)
3. Return only the category label

Text to analyze: {*Text*}
Quadruple: {*Personl, Person2, Time, Location*}
Political party information: {*Person1:Party1, Person2:Party2*}

Figure 8: The prompt for GPT-4o-mini to categorize political interactions.

A.7 Network Construction and Polarization Metrics

A.7.1 Weighting Scheme

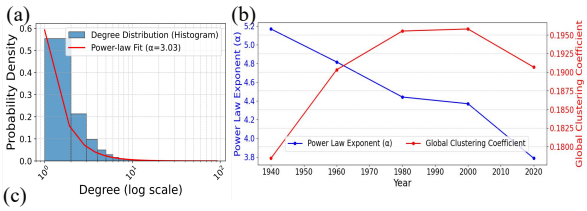
In our network analysis, nodes represent political figures, and edges represent interactions. Following the Weighted Signed Networks (WSN) framework (Kumar et al., 2016), we assigned specific signed weights to capture both polarity and magnitude:

- -2 (*Adversarial*): Reflects strong negative relational dynamics.
- $+2$ (*Cooperative*): Reflects strong positive collaboration.
- $+1$ (*Neutral*): Assigned a positive but lower weight. Unlike previous studies that might assign 0, we argue that neutral interactions in political contexts imply a baseline of civil discourse and goodwill, distinct from the absence of interaction.

A.7.2 Calculation of Standardized Modularity

Following Conover et al. (Conover et al., 2011), we implement their normalized modularity calculation method to quantify polarization. As direct comparison of modularity values across networks with varying sizes and connection densities is challenging, their approach evaluates the relative quality of cluster assignments. The implementation involves:

- Generating $N = 1000$ randomized network samples preserving both degree sequence and edge weight distribution (for weighted graphs in our scenario).



Time Period	Artist	Scientist	Athlete
1920 - 1940	Fletcher Henderson	Albert Einstein	Babe Ruth
1940 - 1960	Frank Sinatra	Sydney Brenner	Antonino Rocca
1960 - 1980	Bob Dylan	Andrei Sakharov	Warren Bockwinkel
1980 - 2000	Robert De Niro	Andy Bechtolsheim	Lindsay Davenport
2000 - 2020	Jennifer Lopez	Stephen Hawking	LeBron James

Figure 9: (a) In the constructed network, the distribution of node degrees adheres to a power-law distribution. (b) Clustering coefficient c and power-law distribution parameter α over 20-year intervals from 1920 to 2020. (c) From 1920 to 2020, every 20 years, the top-ranked individuals in the professions of artists, scientists, and athletes according to their PageRank.

- Computing the mean (μ) and standard deviation (σ) of modularity Q from randomized samples.
- Deriving the Z-score: $Z = (Q_{\text{original}} - \mu)/\sigma$.

This normalization procedure quantifies how significantly the observed community structure deviates from random expectation, while maintaining the original network's topological characteristics through degree- and weight-preserving randomization.

A.8 Additional Analysis

A.8.1 Overall Analysis

Figure 10 illustrates the global distributions of these interactions, showing that most interactions occur in North America and Europe, followed by Southeast Asia and Australia.

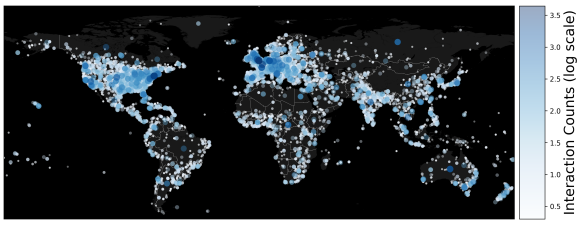


Figure 10: A geographic heatmap of interaction locations, where darker colors indicate a higher number of interaction events.

With these interactions, we then build an interaction network. In Figure 9 (a), we illustrate the degree distribution across the entire network, ob-

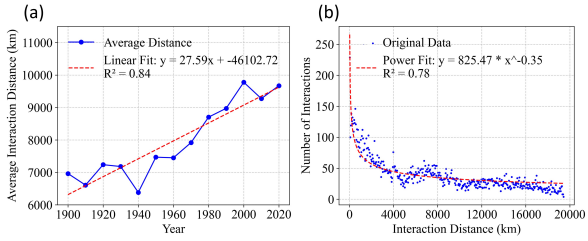


Figure 11: (a) From 1900 to 2020, the evolution of the average interaction distance. (b) The distributional relationship between the number of interactions and the distance of these interactions.

serving a conforming pattern to the power-law distribution.

We further examine the temporal perspective. Using data from each 20-year interval from 1920 to 2020, we construct 5 sub-networks. For each sub-network, we calculate the clustering coefficient c and the power-law distribution parameter α. As we can see from Figure 9 (b), c exhibits an increasing trend, while the α shows a decreasing trend, indicating the interaction networks are becoming centralized over time.

For each 20-year interval, we identify the top individuals in the occupations of artist, scientist, and athlete, respectively, ranked by PageRank scores. This is illustrated in Figure 9 (c).

A.8.2 Interaction Distance

We define interaction distance as the cumulative distance from *Location* to both *Person1*'s and *Person2*'s birthplaces, indicating the travel distance required for the interaction. As illustrated in Figure 11 (a), the average interaction distance exhibits a growing trend over time, likely attributable to advancements in modern transportation and communication. In contrast, Figure 11 (b) shows that most interactions occur over short distances, with the distribution conforming to a power-law pattern. These findings are consistent with those reported by Illenberger et al. (2013).

A.9 Extraction from News Archives

Building upon the experimental results in Section 6.4, which demonstrated FALCON's robust cross-domain applicability on news media, we extended our extraction pipeline to the large-scale *all-the-news-2-1* dataset[11]. This corpus contains a total of **2,688,868** articles published by major American news outlets between 2016 and 2020.

The extraction process was conducted in two phases:

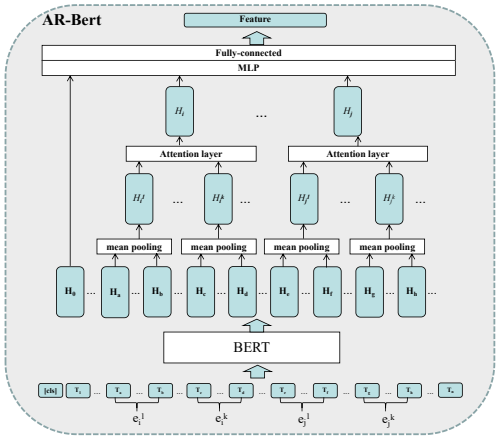


Figure 12: The architecture of AR-BERT.

Candidate Generation. We applied the same trajectory mining pipeline used for Wikipedia to the raw news texts. By filtering for text segments containing the co-occurrence of at least two *Person* entities alongside explicit *Time* and *Location* mentions, we generated **552,534** candidate interaction quadruplets.

Inference and Extraction. We then utilized the FALCON model (trained on Wikipedia) to classify these candidates. Consistent with its performance in our experimental evaluation, FALCON successfully identified **77,836** valid spatio-temporal interaction quadruplets. This resulting dataset captures contemporary interactions spanning from 2016 to 2020.

A.10 AR-BERT Figure

Figure 12 presents a schematic illustration of the AR-BERT framework.

[11] https://huggingface.co/datasets/rjac/all-the-news-2-1-Component-one