arXiv:2507.21903v2 [cs.SI] 31 Jul 2025

Who's important?— SUnSET*: Synergistic Understanding of Stakeholder, Events and Time for Timeline Generation*

Tiviatis Sim[1,2], Yang Kaiwen, Shen Xin, Kenji Kawaguchi[1]
[1]National University of Singapore, [2]A*STAR Institute of High Performance Computing
tiviatis@u.nus.edu, kaiwen_yang@alumni.brown.edu, sxstar@zju.edu.cn, kenji@comp.nus.edu.sg,

Abstract

As news reporting becomes increasingly global and decentralized online, tracking related events across multiple sources presents significant challenges. Existing news summarization methods typically utilizes Large Language Models and Graphical methods on article-based summaries. However, this is not effective since it only considers the textual content of similarly dated articles to understand the gist of the event. To counteract the lack of analysis on the parties involved, it is essential to come up with a novel framework to gauge the importance of stakeholders and the connection of related events through the relevant entities involved. Therefore, we present SUnSET: Synergistic Understanding of Stakeholder, Events and Time for the task of Timeline Summarization (TLS). We leverage powerful Large Language Models (LLMs) to build SET triplets and introduced the use of stakeholder-based ranking to construct a *Relevancy* metric, which can be extended into general situations. Our experimental results outperform all prior baselines and emerged as the new State-of-the-Art, highlighting the impact of stakeholder information within news article.

1 Introduction

The abundance of online news media in the advent of the information era has led to a growing volume of daily data production [17], which poses significant challenges in efficiently identifying and understanding information concisely [1]. To promote efficacy and prevent information overloading, automated tasks such as Timeline Summarization (TLS) [4, 8, 6] becomes especially critical in this era [12, 23]. TLS is the generation of a summarized timeline of events, where multiple date-event pairs are sequentially listed to form a narrative for a particular topic. This process generally involves

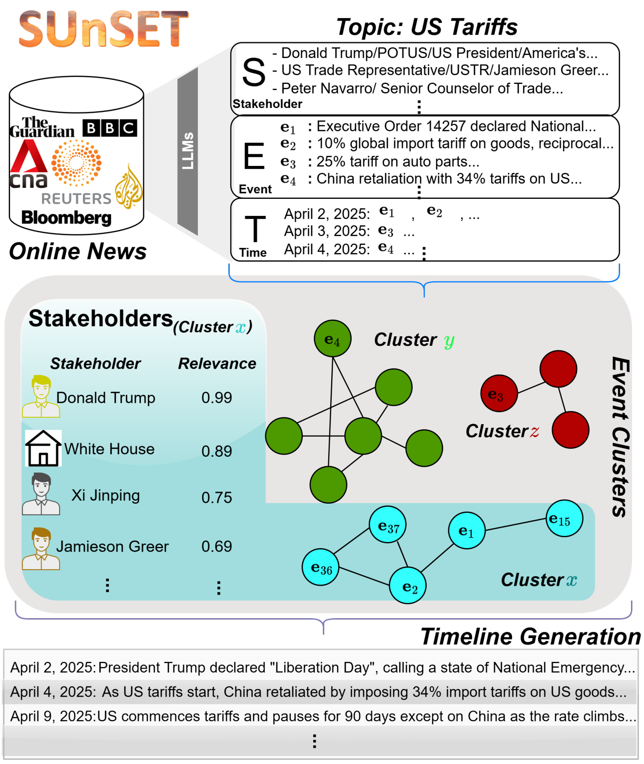

Figure 1: An illustration of how SunSET generates a timeline for TLS through utilizing stakeholder information for relevance scoring.

two essential requirements; (1) the events generated must be sequential and succinct (2) only relevant, fundamental and essential events should be included.

Several papers have attempted to improve the brevity of the first requirement and the identification of the second through the use of powerful Large Language Models (LLMs). These methods typically yield good results [22, 16, 6, 12, 23] due to the emergence of temporal understanding capabilities found within LLMs [24]. Nevertheless, the utilization of LLMs in TLS is still relatively underexplored [6, 16] as compared to works on temporal reasoning such as temporal question answering tasks [11].

Although previous works have utilized LLMs to aid temporal understanding within TLS frameworks, they typically focus on article summarization. Moreover, earlier approaches predating LLMs primarily addressed causal and temporal influences. Yet, stakeholders directly involved in events are crucial for identifying links between events and assessing their significance—an aspect largely overlooked in current TLS research. Incorporating stakeholder information can enhance both sequential curation and importance selection by grounding the timeline in real-world relevance.

As such, we propose SUnSET: Synergistic Understanding of Stakeholder, Events and Time, for the task of automated TLS generation (Fig. 1). Unlike prior methods, SUnSET attempts to extract relevant information of various stakeholders mentioned within articles. Moreover, instead of summarizing articles immediately [12, 6], SUnSET attempts to extract multiple relevant events mentioned within an article to enable a better representation of occurrences documented by a single news article. The extracted stakeholder and event information will subsequently be combined with its estimated timeframe to generate a Stakeholder-Event-Time triplet (SET). This will be utilized in various clustering steps of the timeline generation process to eventually generate a singular summarized timeline. SUnSET was tested against classic TLS datasets [19, 20] and was able to improve existing baselines to emerge as the new state of the art.

Henceforth, the key contributions from our paper include the following:

1. A novel framework is introduced to construct and apply SET for TLS.

2. This is the first work to propose methods for extracting and utilizing stakeholder information within TLS.

3. Multiple scoring mechanisms are developed, along with constraint-based proofs of stakeholder relevance applicable to general cases.

4. Empirical evaluations across TLS datasets demonstrate improved performance, validating the benefits of SUnSET in TLS tasks.

2 Related Works

Recent TLS approaches can be broadly categorized into two main types: traditional methods and LLM-augmented methods; although variants of traditional clustering methods are still popular approaches when it comes to TLS task [5, 3], the rise of emergent capabilities in generative LLMs has led to increasing efforts to leverage these tools for TLS tasks, often with diverse objectives. Although LLM applications are more commonly associated with temporal reasoning, their usage in TLS has grown significantly over the past three years. For instance, Wu et al.'s paper highlighted the importance of temporal and causal relationships by incorporating LLMs into self-questioning strategies. On the other hand, works by Hu et al. and Qorib et al. introduced the notion of LLMs functioning as pseudo-oracles, emulating crowdsourced event clustering via pairwise querying during the clustering process.

Clustering methods in TLS are typically implemented in two distinct ways [5]. The more conventional approach involves grouping key dates to generate graph-based rankings [18] where subsequent summaries will be formed by selecting the top candidate sentences derived from generated ranked lists of important dates. Alternatively, the other method involves 'event' clustering [14], where article-level summaries generated once from each article are aggregated based on textual similarity. Increasingly, LLMs are integrated into these processes, especially to enhance summarization. For example, Hu et al. applied LLMs for article-level summarization prior to clustering, followed by a reclustering stage where the LLM validates cluster cohesion through pairwise comparisons of summary nodes. Meanwhile, Wu et al. explored iterative prompting techniques to survey context by progressively self-refining queries with LLMs to retrieve more relevant content. Building upon these efforts, our paper advances TLS clustering with LLM augmentation, where we will expand on stakeholder-based heuristics to assess event relevance.

3 Methodology

We introduce **SUnSET: Synergistic Understanding of Stakeholder, Events and Time** in this paper for the task of TLS. SUn-

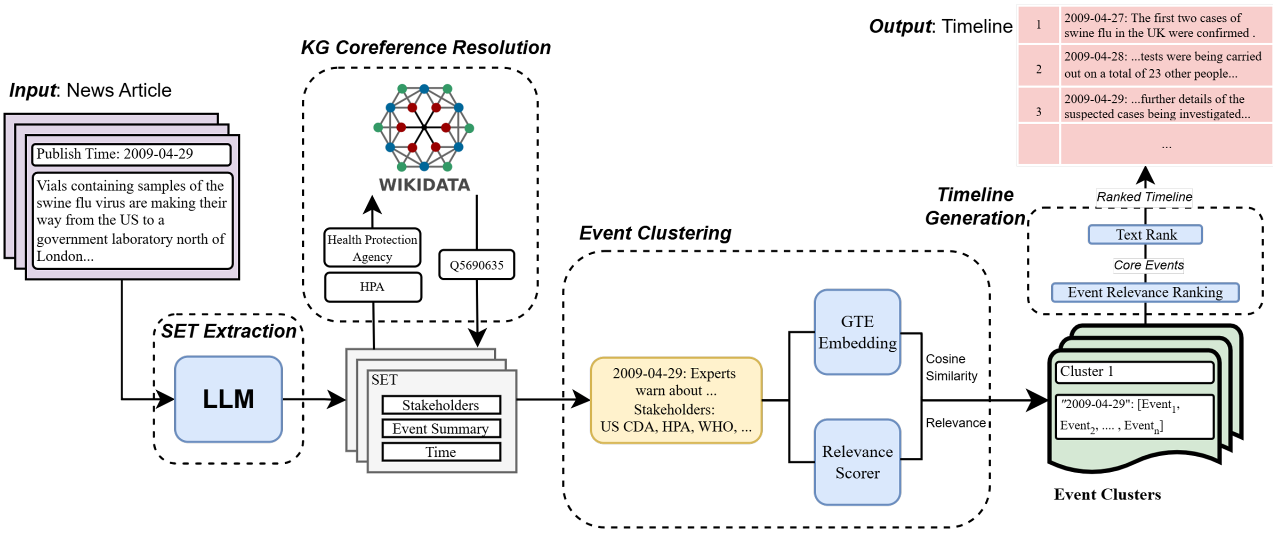

Figure 2: Full SUnSET Framework for TLS. News articles generate SETs and Stakeholders undergo Coreference Resolution. Subsequently, events are clustered through Cosine Similarity and Relevance. The clusters will then be ranked while TextRank extracts the narrative for final timeline creation.

SET aims to utilize stakeholder information while identifying multiple event occurrences within a single article for creating a graph representation of important events interconnected to each other. The general pipeline can be seen in Figure 2, where the pipeline is split into three main sequences.

3.1 SET Generation

The first step of SUnSET requires the generation of Stakeholder, Event and Time to form a triplet (SET). This is done via utilizing a LLM to extract events mentioned within an article and their estimated date of occurrence. Therefore, every article A may contain multiple events: $A \rightarrow \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots$. After extracting the event and its estimated date (t), we call the same LLM to extract relevant stakeholders mentioned within the article with a maximum of 5 stakeholders per event: $\mathcal{S}_\mathbf{e} = (\varsigma_1, \varsigma_2, \varsigma_3, \varsigma_4, \varsigma_5)$. The term stakeholder used here strictly refers to an entity which is either a person or an organization. This results in a series of SET for a single article: $A \rightarrow ((\mathcal{S}_{\mathbf{e}1}, \mathbf{e}_1, t_{\mathbf{e}1}), (\mathcal{S}_{\mathbf{e}2}, \mathbf{e}2, t_{\mathbf{e}2}), (\mathcal{S}_{\mathbf{e}3}, \mathbf{e}3, t_{\mathbf{e}3}), \ldots)$. The relevant prompts used can be found in Appendix A.

Subsequently, Coreference Resolution will be done for all of the extracted stakeholders due to difference in naming such as utilizing a title or a position than a name (E.g. President of the United States) or differences in naming (E.g. POTUS v.s. President of America). A knowledge graph (KG) was built from the wikidata [21] database. The specifics of the KG building can be accessed in Appendix B.

3.2 Event Clustering

Reoccurrence \ Rarity Across Topics	*Rare*	*Common*
Low	Normal	Irrelevant
High	Significant	Normal

Table 1: Matrix showing Stakeholder frequency and relevance; Reoccurrence refers to repeated occurrence within the current topic, whereas rarity looks into frequency of a stakeholder across all topics.

The next stage of the workflow involves the event clustering process. Typically, this involves using content found within similarly dated articles to generate clusters, where each article forms an event node. Instead of using the first five sentences in the body of each article [5] or a pre-generated article summary [6, 12] as the input, SUnSET uses every single event ($\mathbf{e}$) to generate a more accurate representation of an event's impact.

Additionally, unlike the aforementioned clustering methods which uses only date or cosine similarity for cluster creation, we came up with a new metric to gauge the importance of an event. This new metric incorporates stakeholder information; intuitively, the more the stakeholders occurs across all topics D, the less important they are. Similarly, if there is frequent reoccurrence of stakeholders within the same topic ($d \in D$), they are more significant (Tab. 1). Therefore, SUnSET clustering introduces the use of Equation 1 to represent the relevance of stakeholders. This score penalizes common stakes across D (Eqn. 2) while rewarding within-topic reoccurrence (Eqn. 3).

$$Rel(\varsigma, d) = \beta(P(\varsigma, d) \cdot R(count(\varsigma_d))) \quad (1)$$

$$P(\varsigma, d) = \frac{s_{\varsigma_D}}{\bar{x}_{\varsigma_D} \cdot \sqrt{|D|}} \times \frac{count(\varsigma_d)}{count(\varsigma_D)} \quad (2)$$

$$R(x) = \frac{e^{x/10} - e^{-x/10}}{e^{x/10} + e^{-x/10}} \quad (3)$$

$$W_{edge} = \sum_{i}^{SET_d} \sum_{j \neq i}^{SET_d} \{Bool_{EM_n}(\varsigma_{\mathbf{e}i}, \varsigma_{\mathbf{e}j}) \cdot [\sum_{\varsigma_i = \varsigma_j} Rel(\varsigma, d) + cos(\mathbf{e}_i, \mathbf{e}_j)]\} \quad (4)$$

The penalty score (P) utilizes the coefficient of variation [2] to grasp the relative dispersion of stakeholder counts with respect to the sample mean across all topics. This value will be max-min normalized, and multiplied by the percentage of occurrence in the current topic. The boundary of P and related proof will be in Appendix C. On the other hand, the reward score (R) uses a dampened hyperbolic tangent score with a factor of $\frac{1}{10}$ to limit the reward value from exploding, where the maximum possible value is capped at 1. The final relevance score thus multiplies the reward with the penalty score, before using a hyperparameter β to tune the significance of the relevance score.

To test the efficacy of the aforementioned penalty score, we introduce another penalty scoring, which is an altered version of the inverse document frequency (IDF) found in BM25 [15]. Unlike Equation 2, Equation 5 directly uses the absence and presence of stakeholder counts across all documents to estimate the degree to penalize; this value however, fails to capture the actual extent of topic-based counts since it does not consider topic-based absence but document-based absence across the full document set. This may lead to edge cases being misrepresented: cases with common rarity and low reoccurrence (*Irrelevant*) may have same penalty output as those with both rare and high reoccurrence (*Significant*) since both cases provides high $|\forall A_\varsigma \in D|$ values. Contrary to using P_{IDF}, the coefficient of variation in the original penalty score represents the spread of ς across all topics, which allows the subsequent penalization of compact ς counts while identifying low and high frequency topics in high-variation ς counts through percentage multiplications. The behaviour of the different penalty and reward scores will be further illustrated in Appendix D.

$$P_{IDF}(\varsigma) = lg(\frac{|\forall A \in D| - |\forall A_\varsigma \in D| + 0.5}{|\forall A_\varsigma \in D| + 0.5}) \quad (5)$$

To increase the strictness of relevant events within the same cluster, we have also experimented with incorporating stakeholder Exact Matching (EM), where every 2 nodes requires at least N-matching unique stakeholders to have an edge. Therefore, the clustering process uses the encoded event summary from a General Text Embedding (GTE) Model [25] to obtain query-based cosine similarity scores combined with relevancy scores to generate the top 20 similar events for every node. These events have to satisfy the boolean EM condition to gain a weighted connection (W_{edge}) with the query node (Eqn. 4). Due to the use of relevancy scores, our method does not need to incorporate resource-intensive pairwise LLM comparisons of events found in the same cluster [6, 12].

3.3 Timeline Generation

After obtaining multiple clusters, each cluster C undergoes a re-ranking based on both their size and the existing relevance of their events' stakeholders (Eqn. 7); instead of incorporating relevance scoring to aid in edge connection (Eqn. 4), the relevance score is applied to determine how significant each cluster is relative to both $|C|$ and *Rel*. Events which are measured as significant will then be passed into TextRank [9] to identify important nodes within existing clusters. The final set of nodes will subsequently be used in the final Timeline Generation (TLG).

$$\mathcal{S}_C = \bigcup_{\mathbf{e} \in C} \{\varsigma \mid \varsigma \in \mathcal{S}_\mathbf{e}\} \quad (6)$$

$$Significance = [1 + ln(|C|)] \cdot \frac{\sum_{\varsigma \in \mathcal{S}_C} Rel(\varsigma, d)}{|\mathcal{S}_C|} \tag{7}$$

4 Experimental Setups

Datasets. We used two renowned TLS datasets, Timeline17 (T17) [19] and Crisis [20]. T17 contains 19 timelines compiled from varying sources of online news sites, spanning 9 major topics from 2005-2013, each with 1-5 ground truth timelines. On the other hand, Crisis has 22 annotated timelines covering 4 critical crisis events, each containing 4-7 ground truth timelines.

Benchmark. We use two recent SOTA papers as our baselines: Hu et al. and Wu et al. are the current best-performing works and have reproducible results on T17 and Crisis, beating prior popular methods such as CLUST [5] and EGC [8].

Models and Deployment. For experiments, we adopted Qwen2.5-72B-Instruct [13], and GPT-4 omni [10]. We used GTE-Modernbert-Base for encoding. For the main experimental results, we used the best model, Qwen2.5-72B-Instruct to compare across the different benchmarks. All model deployment were done with VLLM [7].

Metrics. Adhering to previous work done [6, 12, 23], we incorporate a part of the Tilse framework [18] and use three main scores to analyse the performance of the TLS task. We use an Alignment-based ROUGE-1 F1-score (AR-1) to evaluate the semantic distance of unigram overlaps between generated timelines and the provided ground truth. We also used the corresponding bigram overlaps and scored the Alignment-based ROUGE-2 F1-score (AR-2). Lastly, the Date-F1 metric will score the similarity of the dates in the generated timeline compared to the referenced ground truth to understand the quality of major events picked up.

Setups. Following prior papers [6, 12], the TLS task can be defined as follows: given a set of temporally labelled news articles that is related to a broad topic ($d \to A_1, A_2, A_3 ..., \forall d \in D$), as well as the expected number of dates and the number of sentences to include in each date, a single timeline summarizing the topic should be generated. This generated timeline will subsequently be compared with the ground truth provided in the datasets. The baselines from the selected benchmarks are run in their publicly available GitHub repository.

We will release our full code publicly on Github once the paper is accepted into the conference.

5 Experimental Results

Table 2 showcases SUnSET as compared to the previously SOTA baselines. We selected the strongest model for each method for the comparisons. We also included Llama2-13B in Table 2 since Llama2 was the original model used in LLM-TLS's paper [6] and it occasionally performs better than Qwen2.5-72B.

Dataset	Method	LLM	AR-1	AR-2	Date-F1
Crisis	CHRONOS	Qwen72B	0.108	0.045	0.323
	LLM-TLS	Llama13B	0.112	0.032	0.329
	LLM-TLS	Qwen72B	0.111	0.036	0.326
	SUnSET	Qwen72B	**0.129**	**0.047**	**0.389**
	SUnSET	GPT-4o	0.107	0.036	0.381
T17	CHRONOS	Qwen72B	0.116	0.042	0.522
	LLM-TLS	Llama13B	0.118	0.036	0.528
	LLM-TLS	Qwen72B	0.114	0.040	0.543
	SUnSET	Qwen72B	**0.136**	**0.044**	0.576
	SUnSET	GPT-4o	0.120	0.039	**0.590**

Table 2: Performance of SUnSET on Crisis and T17 datasets

Our method is able to defeat all prior work in all our metrics used for evaluation. For instance, SUnSET managed to improve Crisis Date-F1 score by 18.2%, ROUGE-1 by 15.2% and ROUGE-2 by 30.6% as compared to the previous SOTA result. The increase in Date-F1 scores over all prior methods indicates the superiority of SUnSET in correctly identifying important dates that were curated in the ground truth, whereas the high ROUGE scores support SUnSET in its ability to retain important details in the distillation process. SUnSET also managed to obtain the best scores in T17, surpassing the previous SOTA. The full set of experimental results are included in Appendix F; examples of timeline output can be accessed in Appendix G. For reference, we attempted using SUnSET with a closed-sourced model (GPT-4omni). GPT-4o performed relatively well and was able to have strong improvements in Date-F1 scores, but lacks in its ROUGE scores as compared to Qwen experiments. To understand where the difference in improvement comes from, more experiments were done on the clustering and timeline generation processes.

5.1 Relevancy in Event Clustering

Table 3: *Rel* in Event Clustering

Dataset	Method	AR-1	AR-2	Date-F1
Crisis	w/o *Rel*	0.114	0.041	0.354
	Rel	**0.129**	**0.047**	**0.389**
T17	w/o *Rel*	0.127	0.038	0.551
	Rel	**0.136**	**0.044**	**0.576**

We further investigated the impact of relevance used in the event clustering process by ablating all stakeholder-related information and relevancy scores, retaining only the cosine similarity scores of the events. From the results observed in Table 3, the use of relevance scoring and stakeholder information causes a leap in performance on Date-F1 scores by 10% for Crisis and 4.5% for T17. This indicates that our method for the identification of stakeholders plays an integral part in ranking event importance. Moreover, even without implementing *Rel*, it is observed that utilizing LLMs to generate event sets $\mathbf{e}$ in SET is already more effective than existing methods which typically uses summarization of A [12, 6] or a self-questioning rewriter [23]. Even without pairwise matching ("w/o *Rel*"), using only the cosine similarity of $\mathbf{e}$ managed to beat the benchmark set by LLM-TLS and CHRONOS (Tab. 2). This reveals that extraction of multiple mentioned events within a single article is vital for TLS tasks; this is intuitive, as news articles often reference related developments that connect to the central event being reported. Significant occurrences may receive limited coverage initially, until a subsequent development—part of an ongoing sequence—draws renewed attention from multiple news outlets.

5.2 Relevancy in Timeline Generation

As stakeholder relevance was utilized in the Timeline Generation process, we compared the difference between the use and absence of *Rel* to estimate the significance of each cluster before passing them into TextRank . We used the best performing results for both settings (Tab. 4). The inclusion of relevance scoring improved the performance across all metrics for both Crisis and T17; this demonstrates the efficacy of utilizing *Rel* for cluster ranking, which ultimately improved the selection of important dates while retaining the details of important events.

Table 4: *Rel* in Timeline Generation

Dataset	Method	AR-1	AR-2	Date-F1
Crisis	TextRank	0.117	0.040	0.368
	TextRank+*Rel*	**0.129**	**0.047**	**0.389**
T17	TextRank	0.128	0.041	0.559
	TextRank+*Rel*	**0.136**	**0.044**	**0.576**

To go a step further, we looked into the influence of the β hyperparameter over the entire SUnSET method, where we compared the inclusion and exclusion of *Rel* during the timeline generation process (Fig. 3). In the full SUnSET process (TextRank+*Rel*), all three scores (ROUGE-F1, ROUGE-F2 and Date-F1) increases as β increases. This behaviour is more latent when *Rel* is only used in clustering (TextRank only) as a subsequent decrease in performance is observed as β increases beyond plateau.

There are several differences in Crisis and T17's increase as β grows. T17, which contains a larger amount of topics but a smaller pool of articles within each topic (larger D and smaller A) tends to have a sharp and significant increase with a small β at 0.1; although it shows the use of *Rel* is effective once introduced, the performance may subsequently become stagnant with less significant changes with the increment of β. In contrary, Crisis with well established topics containing large pools of articles tend to benefit more as β increases after its initial introduction, where small β values are not helpful and potentially worsens performance, and larger β values provide remarkable growth.

5.3 Penalty PvP: P and P_{IDF}

As we introduced a total of two different penalty scores, we tested the difference in the type of penalty used within the relevance score for both datasets. Table 5 shows the difference in using P and P_{IDF} for SUnSET, where the best result out of all β attempted was used.

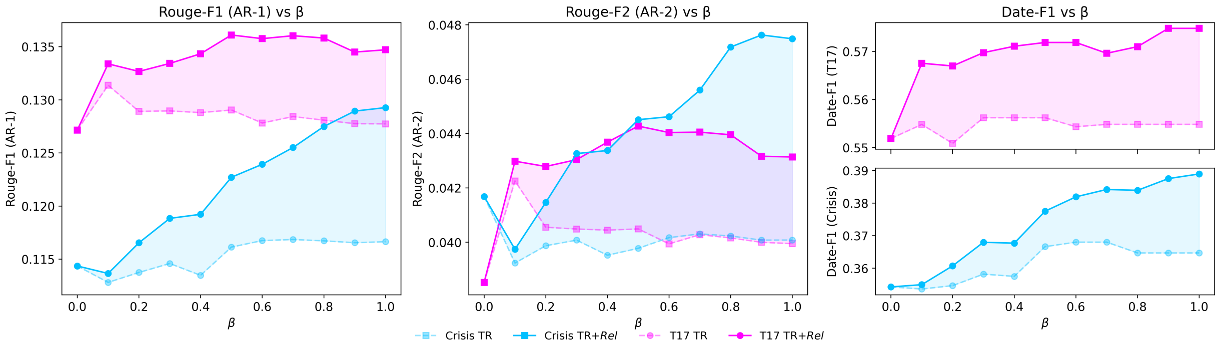

Figure 3: Effect of β hyperparameter on *Rel* in event clustering (TR) and *Rel* in both event clustering and timeline generation (TR+*Rel*). The leftmost graph compares ROUGE-F1 values, the middle graph compares ROUGE-F2 values, and the rightmost graphs compare Date-F1 values.

Table 5: Penalty PvP- P versus P_{IDF}

Dataset	Penalty	AR-1	AR-2	Date-F1
Crisis	P_{IDF}	0.119	0.043	0.374
	P	**0.129**	**0.047**	**0.389**
T17	P_{IDF}	**0.137**	0.043	**0.576**
	P	0.136	**0.044**	**0.576**

We observed that the impact on T17, which contains smaller datasets and thus lesser stakeholders than Crisis, is smaller; the use of either P and P_{IDF} causes little differences in T17's final results, yet it causes significant differences in a larger dataset such as Crisis. This indicates that P is more flexible in more cases, especially when there exist a larger event space where more *Significant* and *Irrelevant* stakeholders (Tab. 1) can be identified.

5.4 Stakeholder Exact Matching

Table 6: Performance of using EM in Event Clustering; best β values are recorded when $\beta > 0$

Dataset	β	EM	AR-1	AR-2	Date-F1
Crisis	0	0	0.114	0.041	0.354
		1	**0.119**	**0.044**	**0.364**
		2	0.114	0.041	0.358
	1.0	0	**0.129**	**0.047**	**0.389**
	0.9	1	0.127	0.046	0.386
	0.6	2	0.121	0.042	0.379
T17	0	0	0.127	0.038	0.553
		1	**0.128**	**0.040**	0.553
		2	0.127	**0.040**	**0.558**
	1.0	0	0.134	0.043	0.574
	0.9	1	**0.136**	**0.044**	**0.576**
	1.0	2	0.134	0.044	0.573

Our initial experiments used only EM with SET without consideration of relevance scores. In Table 6, when β is set as 0, there is no relevance scoring used, and we observe that the inclusion of EM is able to capture meaningful stakeholder information to a proportional degree. Typically, EM_1 is sufficient to result in an improvement, and further matching counts may instead deteriorate the performance of models. This may be due to insufficient stakeholders with high counts, where the median of repetitive stakeholders typically lies around 3 across all datasets (Appn. E). Though EM performance does not match up to the addition of *Rel*, this is sufficient to signify the importance of stakeholder information in the TLS.

Furthermore, the improvement of *Rel* over EM illustrates the strength of our formulated equation in stakeholder ranking. While *Rel* is used, the use of EM causes limited improvements as *Rel* in itself is able to capture and rank the entities involved in **e**. It is also observed that high β values typically performs better across differing EM settings, which further emphasizes the importance of *Rel*.

5.5 Time Comparison

Table 7: GPU hours for clustering

Dataset	Method	GPU Time
Crisis	LLM-TLS	7 hours 12 min
	CHRONOS	24 min
	SUnSET	7 ***min*** 14 sec
T17	LLM-TLS	2 hours 12 min
	CHRONOS	1 hour 9 min
	SUnSET	1 ***min*** 41 sec

Lastly, we measured the GPU hours required for running the experiments, and realised the high efficiency for using SUnSET as compared to previous SOTA methods. The benefits mainly comes from the lack of additional LLM prompting beyond SET creation, where CHRONOS utilizes questioning rounds and LLM-TLS prompts once for every pairwise comparison required. Therefore, SUnSET is able to bypass such time-consuming prompt-questioning methods due to the high efficacy of utilizing stakeholder information for distilling events into a compact representation.

6 Conclusions

We propose SUnSET, a novel approach that leverages SET triplets and stakeholder relevance to generate milestone events for TLS problems. By incorporating stakeholder-relevancy heuristics, SUnSET effectively addresses the challenges of distilling multiple event sets from complex articles. Our framework systematically tackles these issues in three critical steps: 1) SET generation expands the pool of candidate events per article; 2) SUnSET's efficient event clustering eliminates the need for time-intensive reclustering, even with larger candidate sets; 3) Stakeholder information encoding via *Rel* enhances both clustering and timeline generation by yielding a more representative set of linked events. Through extensive experiments and ablation studies on the T17 and Crisis datasets, SUnSET consistently delivers state-of-the-art results, validating its effectiveness in both performance and interpretability. These findings underscore the potential of stakeholder-aware TLS systems to advance narrative coherence, relevance, and human-aligned summarization across evolving news domains, while also establishing stakeholder scoring as a scalable framework for prioritizing information in other contexts where audience-centric relevance and societal impact are key — such as search, recommendation, and policy communication.

Limitations

Our study advances TLS through the integration of SET and stakeholder ranking mechanisms, though certain limitations remain. Within our domain of control, we applied feasible refinement techniques such as prompt engineering and performed rigorous reviews to minimize inconsistencies. However, as with prior work, we were unable to fully mitigate hallucinations or factual inaccuracies that may arise during event extraction when utilizing LLMs due to their inherent generation behavior. Additionally, our experiments were constrained to two representative LLMs, one open-source and one closed-source model, which limits broader generalization across models. Finally, in line with previous studies, we evaluated SUnSET using publicly available research datasets, without extending analysis to real-time news data with more dynamic, evolving topics. We identify this as a promising direction for future exploration.

Acknowledgments

We would like to thank Austen Jeremy Sugiarto and Leong Chee Kai in aiding the initial stages of the experiments. We are also grateful to Dr. Basura Fernando for his help in reading through the paper. We deeply appreciate any reviewers for their helpful feedback.

References

[1] M. Arnold, M. Goldschmitt, and T. Rigotti. Dealing with information overload: A comprehensive review. *Frontiers in Psychology*, 14, Jun 2023. doi: 10.3389/fpsyg.2023.1122200.

[2] C. E. Brown. *Coefficient of Variation*, pages 155–157. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998. ISBN 978-3-642-80328-4. doi: 10.1007/978-3-642-80328-4_13. URL https://doi.org/10.1007/978-3-642-80328-4_13.

[3] X. Chen, M. Li, S. Gao, Z. Chan, D. Zhao, X. Gao, X. Zhang, and R. Yan. Follow the timeline! generating abstractive and extractive timeline summary in chronological order, 2023. URL https://arxiv.org/abs/2301.00867.

[4] H. L. Chieu and Y. K. Lee. Query based event extraction along a timeline. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, page 425–432, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138814. doi: 10.1145/1008992.1009065. URL https://doi.org/10.1145/1008992.1009065.

[5] D. Gholipour Ghalandari and G. Ifrim. Examining the state-of-the-art in news timeline summarization. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 1322–1334, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.122. URL https://aclanthology.org/2020.acl-main.122/.

[6] Q. Hu, G. Moon, and H. T. Ng. From moments to milestones: Incremental timeline summarization leveraging large language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7232–7246, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.390. URL https://aclanthology.org/2024.acl-long.390/.

[7] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[8] M. Li, T. Ma, M. Yu, L. Wu, T. Gao, H. Ji, and K. McKeown. Timeline summarization based on event graph compression via time-aware optimal transport. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6443–6456, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.519. URL https://aclanthology.org/2021.emnlp-main.519/.

[9] R. Mihalcea and P. Tarau. TextRank: Bringing order into text. In D. Lin and D. Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3252/.

[10] OpenAI et al. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

[11] B. Piryani, A. Abdallah, J. Mozafari, A. Anand, and A. Jatowt. It's high time: A survey of temporal information retrieval and question answering, 2025. URL https://arxiv.org/abs/2505.20243.

[12] M. R. Qorib, Q. Hu, and H. T. Ng. Just what you desire: Constrained timeline summarization with self-reflection for enhanced relevance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):25065–25073, Apr. 2025. doi: 10.1609/aaai.v39i23.34691. URL https://ojs.aaai.org/index.php/AAAI/article/view/34691.

[13] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

[14] S. Ribeiro, O. Ferret, and X. Tannier. Unsupervised event clustering and aggregation from newswire and web articles. In O. Popescu and C. Strapparava, editors, *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 62–67, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4211. URL https://aclanthology.org/W17-4211/.

[15] S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

[16] D. Sojitra, R. Jain, S. Saha, A. Jatowt, and M. Gupta. Timeline summarization in the era of llms. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2657–2661, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657899. URL https://doi.org/10.1145/3626772.3657899.

[17] Statista. News and magazines - worldwide: Statista market forecast. URL https://www.statista.com/outlook/amo/app/news-magazines/worldwide.

[18] J. Steen and K. Markert. Abstractive timeline summarization. In L. Wang, J. C. K. Cheung, G. Carenini, and F. Liu, editors, *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 21–31, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5403. URL https://aclanthology.org/D19-5403/.

[19] G. Tran, T. Tran, N.-K. Tran, M. Alrifai, and N. Kanhabua. *Leveraging Learning To Rank in an Optimization Framework for Timeline Summarization.* 2013.

[20] G. Tran, M. Alrifai, and E. Herder. Timeline summarization from relevant headlines. In A. Hanbury, G. Kazai, A. Rauber, and N. Fuhr, editors, *Advances in Information Retrieval*, pages 245–256, Cham, 2015. Springer International Publishing. ISBN 978-3-319-16354-3.

[21] D. Vrandečić and M. Krötzsch. Wikidata. *Communications of the ACM*, 57(10):78–85, Sep 2014. doi: 10.1145/2629489.

[22] S. Wang, Y. Li, H. Xiao, L. Deng, and Y. Dong. Web news timeline generation with extended task prompting, 2023. URL https://arxiv.org/abs/2311.11652.

[23] W. Wu, S. Huang, Y. Jiang, P. Xie, F. Huang, and H. Zhao. Unfolding the headline: Iterative self-questioning for news retrieval and timeline summarization. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4385–4398, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.248. URL https://aclanthology.org/2025.findings-naacl.248/.

[24] S. Xiong, A. Payani, R. Kompella, and F. Fekri. Large language models can learn temporal reasoning, 2024. URL https://arxiv.org/abs/2401.06853.

[25] X. Zhang, Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang, M. Zhang, W. Li, and M. Zhang. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval, 2024. URL https://arxiv.org/abs/2407.19669.

A Prompts used for LLMs

Prompt for Event and Time Generation

```
You are a professional journalist that is tasked to generate date-based
event summary of a given article. A single list contains an article and
its published time. You should generate a dictionary of the most
relevant events of an article, where each key in the dictionary is a
string of the expected event start date in terms of Year-Month-Day
(e.g.2011-12-25) and the value will be a summary of the relevant events
on that day. Summarize only the most important events found in the
article, as succinctly as possible. If you are uncertain of the date of
an event, feel free to use the published date. You should only output
the dictionary in your answer. Generate a dictionary of events of the
following article: {str(article_x)}.
```

Prompt for Stakeholder Generation

```
You are a professional journalist that is tasked to generate the most
relevant stakeholders relevant to a given event summary of an article.
A single list contains an article and its published time. You should
generate a singular list containing not more than five relevant
stakeholders related to only the stipulated event mentioned. These
stakeholders should not be general, and must be identifiable named
entities that can be matched to a person, organization or role when read
on its own. Every single stakeholder generated should also ideally exist
in exact wording as mentioned within the original article. You should
only output the list of stakeholders in your answer, and all
stakeholders should be enclosed in string format. Generate a list of
related stakeholders of event: {dict[key_x]}.
Given article: {str(article_x)}.
```

B Building Knowledge Graph from Wikidata[1]

Algorithm 1 Knowledge Graph for Coreference Resolution

Initialize an empty dictionary: $d \leftarrow \{\}$
for Every stakeholder ς in $\mathcal{S}$ **do**
    if ς in d **then**
        continue
    else
        Search ς in Wikidata label/alt-label with API
        if output O does not exist **then**
            $\varsigma_\gamma \leftarrow$ Remove title in ς with NER
            Search ς_γ in Wikidata label/alt-label with API
            if output O does not exist **then**
                $\varsigma_\delta \leftarrow$ Remove whitespace and replace with `&&`
                Search ς_γ in Wikidata label/alt-label with API
                if output O does not exist **then**
                    Use request API for Wikidata interface search
                    if output O does not exist **then**
                        $d[\varsigma] \leftarrow \varsigma$
                        continue
                    end if
                end if
            end if
        end if
    end if
    if Operator "Position Held By" exist in O **then**
        $P \leftarrow$ ID under "Position Held By"
        $d[\varsigma] \leftarrow P$
        continue
    else
        $d[\varsigma] \leftarrow O$
        continue
    end if
end for

[1]Main API: https://www.wikidata.org/wiki/Wikidata:REST_API, NER module: https://huggingface.co/spacy/en_core_web_trf, Request module: https://requests.readthedocs.io/en/latest/

C Boundary for Penalty Score

Proof. Let $x \in \mathbb{Z}$ and $x \geq 0$ since x represents counts. $|D|$ represents the size of all topics.

$$CV = \frac{s}{\bar{x}} \tag{8}$$

From Equation (8),

$$CV^2 = \frac{\frac{\sum(x_i - \bar{x})^2}{|D|-1}}{\bar{x}^2} = \frac{\frac{\sum x_i^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2}{|D|-1}}{\bar{x}^2} = \frac{\frac{\sum x_i^2 - 2|D|\bar{x}^2 + |D|\bar{x}^2}{|D|-1}}{\bar{x}^2} \tag{9}$$

Since n>1 and $\bar{x} \geq 0$,

$$\sum_{i=1}^{|D|} x_i^2 \leq (\sum_{i=1}^{|D|} x_i)^2 = |D|^2\bar{x}^2 \tag{10}$$

From Equation (10), minus $\bar{x}^2|D|$,

$$\sum_{i=1}^{|D|} x_i^2 - |D|\bar{x}^2 \leq |D|^2\bar{x}^2 - |D|\bar{x}^2 \tag{11}$$

Change Equation (11) LHS to Equation (9),

$$CV^2 = \frac{\frac{\sum x_i^2 - |D|\bar{x}^2}{|D|-1}}{\bar{x}^2} \leq \frac{\frac{|D|^2\bar{x}^2 - |D|\bar{x}^2}{|D|-1}}{\bar{x}^2} = |D| \tag{12}$$

From Equation (12), since $n > 1$ and $\bar{x} \geq 0$,

$$0 \leq CV \leq \sqrt{|D|}$$

From Equation (2), we divided CV by $\sqrt{D}$, and multiplied it by the percentage of ς belonging in d, where $0 \leq \frac{count(\varsigma_d)}{count(\varsigma_D)} \leq 1$, therefore,

$$0 \leq P \leq 1$$

□

D Behaviour of Reward and Penalty Scores

We illustrate the behaviour of both Penalty and Reward in this section. From Figure 4, Graph 1 plots R (Eqn. 3) as the count of stakeholder increases. It can be observed that when the count of the stakeholders ($x \leftarrow count(\varsigma_d)$) reach approximately 21, the value of R saturates regardless of any more addition to counts for the same stakeholder. Subsequently, Graph 2 plots the equation of unnormalized P (Eqn. 2) of 12 different scenarios (Tab. 8).

Index	**Stakeholder Counts**	**Rationale**
0	A: 2, B: 3, C:5	Close Distribution, All Low Counts
1	A: 90, B: 85, C:65	Close Distribution, All High Counts
2	A: 5, B: 5, C:5	Uniform Distribution, All Low Counts
3	A: 16, B: 16, C:16	Uniform Distribution, All High Counts
4	A: 15, B: 4, C:54	Single-Peak Distribution, Peak High Count
5	A: 1, B: 8, C:2	Single-Peak Distribution, Peak Low Counts
6	A: 6, B: 7, C:1	Double-Peaks Distribution, Peaks Low Counts
7	A: 21, B: 19, C:3	Double-Peaks Distribution, Peaks High Counts
8	A: 3, B: 0, C:0	Single-Peak Distribution, Peak Low Count, Rest are 0
9	A: 19, B: 0, C:0	Single-Peak Distribution, Peak High Count, Rest are 0
10	A: 6, B: 3, C:3	Single-Peak Distribution, Peak Low Count, Rest are half
11	A: 26, B: 13, C:13	Single-Peak Distribution, Peak High Count, Rest are half

Table 8: Scenarios used for Stakeholder Counts across Three Topics (A, B and C)

Graph 3 plots the value of Rel without consideration of β (by setting $\beta = 1$) to understand the impact of multiplying P and R. Cases 0, 1, 2, 3, 10 and 11 illustrates when stakeholder counts are well spread across the topics, their P values are low irregardless of high or low counts, which is the desired behaviour since stakeholder with such behaviours fall under either 'Normal' or 'Irrelevant' (Tab. 1). Alternatively, 'Significant' cases with strong single-peak distribution dominating over the others (Indices 4, 5, 8, 9) show significantly higher P values, where completely unique stakeholders tend to reach maximum possible P. It is noteworthy to add that the difference between the final Rel score between cases 8 and 9 is important since the R scores are able to dampen case 8's score due to its low count value which was set to the common median (3) for illustration. This also occurs between cases 4 and 5. Lastly, those with double-peak distribution (Indices 6 and 7) have a mid-low P value, where the duo peaks' P are significantly higher than that of the topic with the low-value count, yet not as high as what would occur in a single-peak. The duo-peaks were penalized with the number of similarly 'Significant' topics within themselves. This leads to a lower overall Rel score.

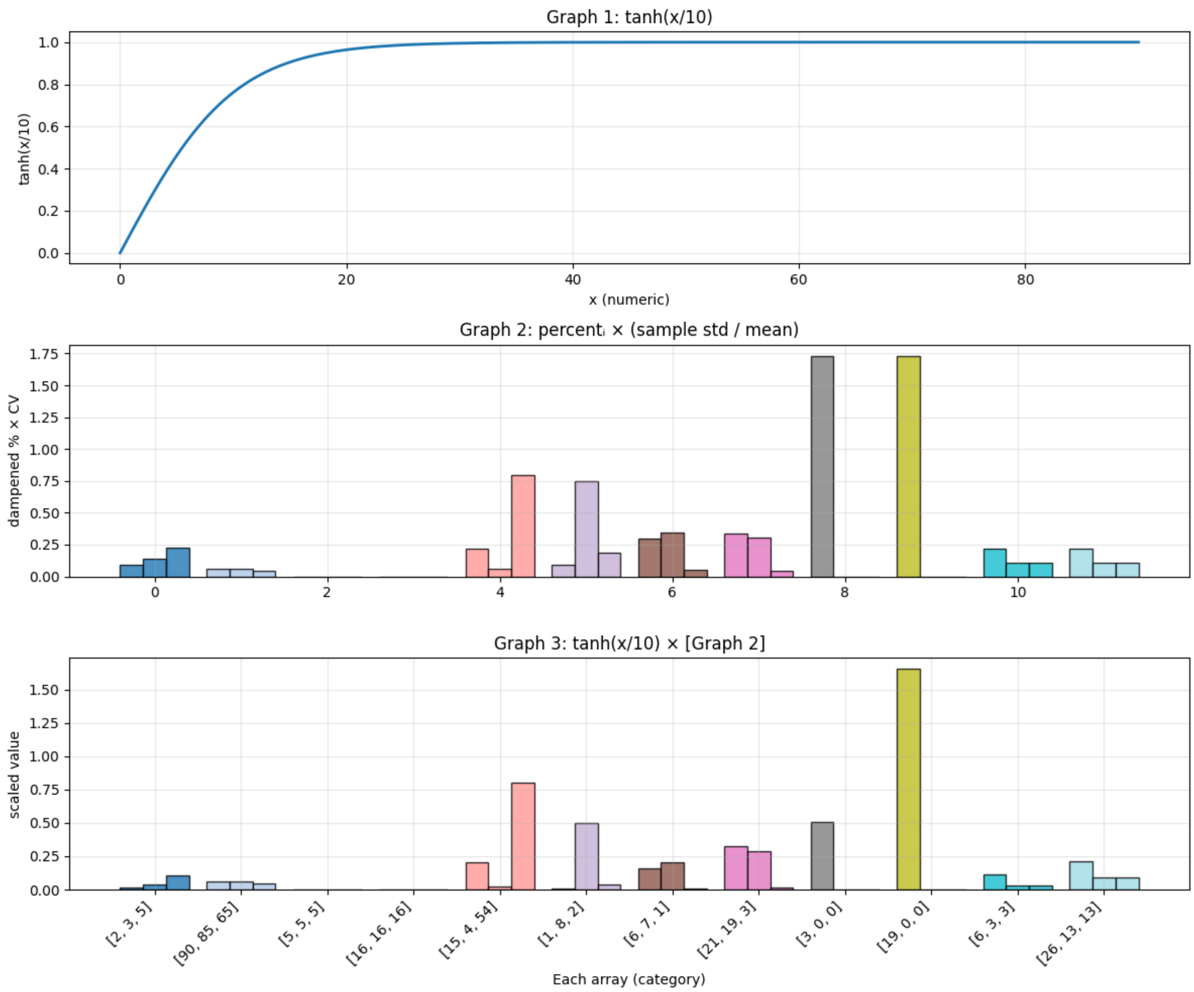

Figure 4: Case-by-case representation of Penalty Behaviour

Similarly, we will examine the behaviour of P_{IDF}. Graph 1 of Figure 5 also shows R for reference. Graph 2 plots unnormalized P_{IDF} value to show how it changes as $|A_\varsigma \in D|$ increases (Eqn. 5). It should be noted that indices 1, 3, 4, 7, 9 and 11 (Tab. 8) cannot be differentiated between topics (i.e. A, B and C) since there is no differentiation between topics while using P_{IDF} which makes it inferior to P. Despite so, it can be observed from Graph 2 that as more articles uses the same stakeholder, P_{IDF} becomes smaller, mirroring the way inverse document frequencies work.

Graph 3 further documents the interaction of R and P_{IDF}, where every line shows how the differing counts of stakeholders across topic ($|A_\varsigma \in D|$) interact when P_{IDF} multiplies R which rewards within a topic ($count(\varsigma_d)$). It is observed that the relevance score of 'Irrelevant' cases (Tab. 1) will still be highly rewarded beyond 1 once $count(\varsigma_d)$ goes beyond 5 regardless of $|A_\varsigma \in D|$, for instance having 5 out of 100 of any stakeholder ς allows a Rel of around 1, which is not ideal despite its proportionally low count value. This flaw from P_{IDF} shows that proportion and distribution should be essential, which is accounted for within P.

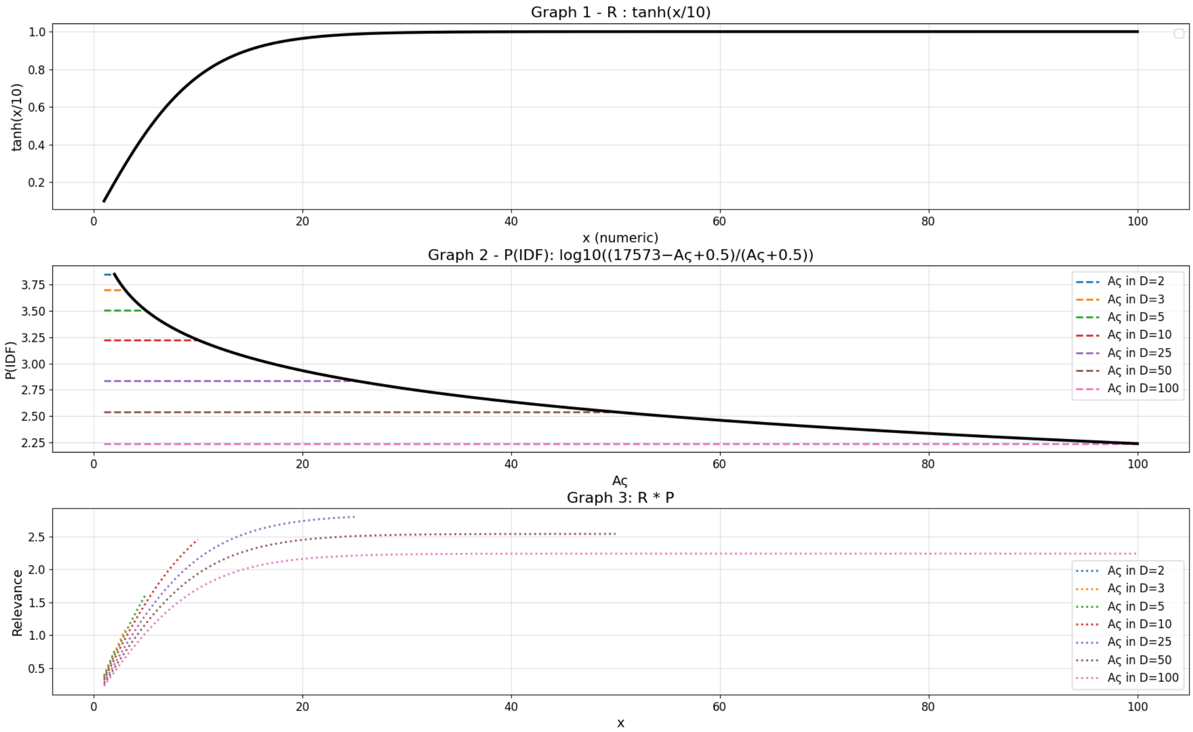

Figure 5: Representation of Penalty$_{IDF}$ Behaviour

E Repeating Stakeholder Counts

We recorded the counts for each stakeholder across all topics in our dataset. There are occasional outliers where the entity involved may be mentioned more than a hundred times and may even reach the thousands, but typically, a count of ten or less is normal across all stakeholders existing in all topics. From Figures 6 and 7, we can observe that the individual median of each topic regardless of the dataset is around three. This means that within their own topics, most stakeholders are mentioned at least in three different articles.

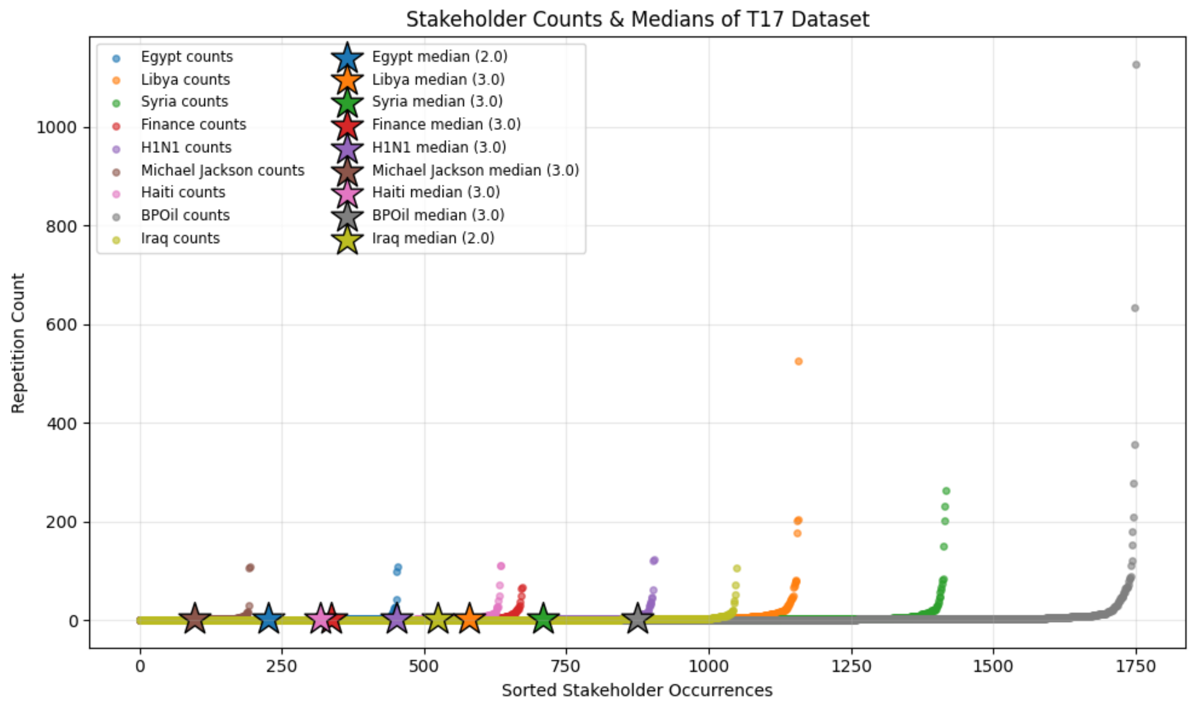

Figure 6: Stakeholder counts across topics within T17 Dataset

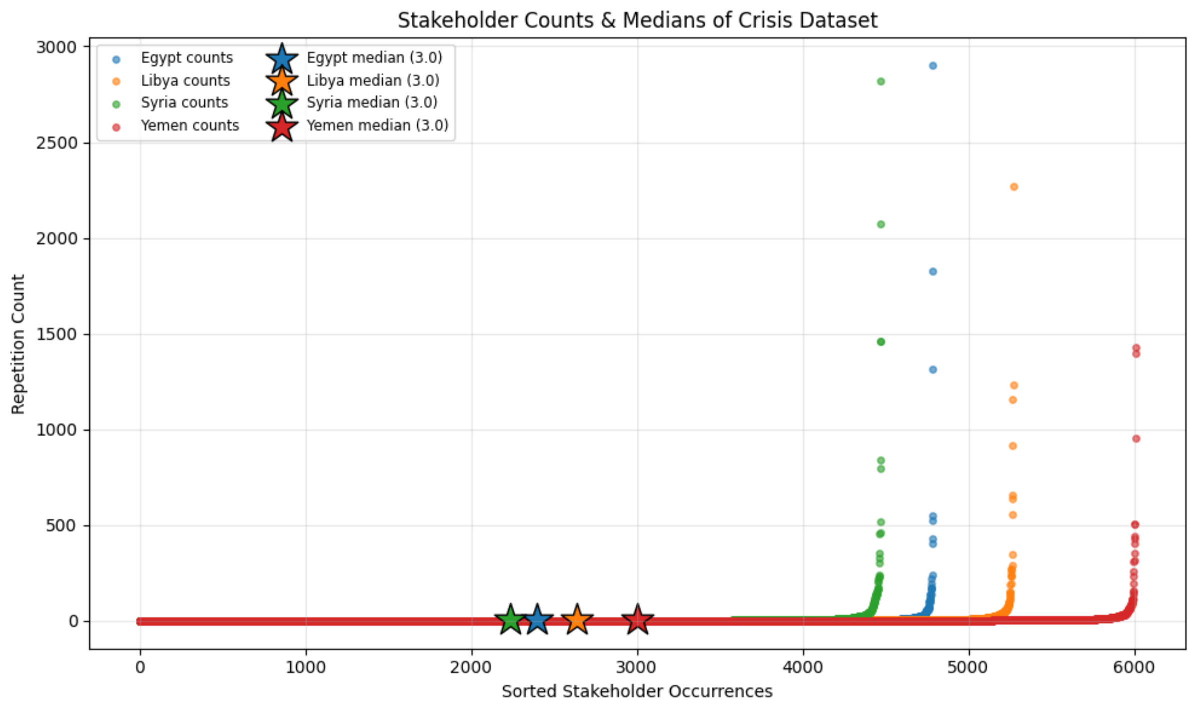

Figure 7: Stakeholder counts across topics within Crisis Dataset

F Full Set of Experiment Results

Table 9: Crisis Results-SUnSET with P for Rel

		TextRank+*Rel*			TextRank		
beta	EM	AR-1	AR-2	Date-F1	AR-1	AR-2	Date-F1
0	0	0.114358	0.041683	0.354251	0.114358	0.041683	0.354251
0.1	0	0.113652	0.039741	0.35494	0.112811	0.039234	0.353606
0.2	0	0.116541	0.041465	0.360721	0.113745	0.039872	0.354645
0.3	0	0.118841	0.043264	0.367969	0.11459	0.040077	0.358176
0.4	0	0.119222	0.043375	0.367698	0.113476	0.039522	0.357489
0.5	0	0.122711	0.04451	0.377522	0.116135	0.039776	0.366623
0.6	0	0.12393	0.04462	0.381933	0.116748	0.040169	0.368
0.7	0	0.125512	0.0456	0.384202	0.116847	0.040308	0.368
0.8	0	0.127489	0.047185	0.383934	0.116723	0.040229	0.364686
0.9	0	0.12894	0.047628	0.387541	0.116548	0.040077	0.364686
1	0	0.129269	0.047488	0.388972	0.116641	0.04008	0.364686
0	1	0.119383	0.044027	0.364554	0.119383	0.044027	0.364554
0.1	1	0.11499	0.040711	0.358591	0.114242	0.040103	0.357267
0.2	1	0.117851	0.04203	0.363621	0.115888	0.041408	0.361055
0.3	1	0.119364	0.042948	0.368201	0.115872	0.041243	0.361055
0.4	1	0.119155	0.042993	0.37158	0.115795	0.04119	0.360607
0.5	1	0.120324	0.043291	0.374345	0.116859	0.040401	0.366181
0.6	1	0.123812	0.044156	0.385318	0.116893	0.040489	0.366181
0.7	1	0.124445	0.044881	0.383322	0.117104	0.040751	0.366181
0.8	1	0.125976	0.04591	0.383054	0.117077	0.040633	0.366181
0.9	1	0.127911	0.046209	0.386313	0.116135	0.039776	0.366623
1	1	0.124948	0.045573	0.374559	0.114044	0.039693	0.366867
0	2	0.114743	0.041232	0.358151	0.114743	0.041232	0.358151
0.1	2	0.114549	0.039838	0.365324	0.113659	0.039767	0.361014
0.2	2	0.115156	0.039829	0.367784	0.113425	0.039519	0.361014
0.3	2	0.118806	0.041571	0.371988	0.114599	0.040124	0.361038
0.4	2	0.120898	0.04296	0.37482	0.115251	0.040805	0.363499
0.5	2	0.121954	0.043179	0.377068	0.115556	0.040812	0.363499
0.6	2	0.121427	0.04293	0.379509	0.116569	0.040968	0.365772
0.7	2	0.120243	0.042156	0.374532	0.1158	0.040556	0.364149
0.8	2	0.120882	0.042673	0.375772	0.115812	0.040559	0.364149
0.9	2	0.122365	0.043452	0.375772	0.115122	0.04015	0.365345
1	2	0.12309	0.043711	0.375196	0.115122	0.040148	0.365345

Table 10: T17 Results-SUnSET with P for *Rel*

		TextRank+*Rel*			TextRank		
beta	EM	AR-1	AR-2	Date-F1	AR-1	AR-2	Date-F1
0	0	0.127151	0.038522	0.551891	0.127151	0.038522	0.551891
0.1	0	0.133391	0.042983	0.567552	0.131371	0.042254	0.554819
0.2	0	0.132674	0.042788	0.566986	0.128913	0.040551	0.550866
0.3	0	0.13343	0.04304	0.569759	0.128958	0.040486	0.556219
0.4	0	0.13434	0.043684	0.571095	0.128798	0.040446	0.556219
0.5	0	0.136108	0.044275	0.571878	0.129042	0.040489	0.556219
0.6	0	0.135763	0.044037	0.571878	0.12782	0.03994	0.55434
0.7	0	0.13604	0.044052	0.569624	0.128425	0.040278	0.554834
0.8	0	0.13582	0.043954	0.571009	0.128086	0.040158	0.554834
0.9	0	0.134498	0.043165	0.574814	0.127756	0.03997	0.554834
1	0	0.134713	0.043138	0.574814	0.127738	0.039948	0.554834
0	1	0.128352	0.040235	0.553264	0.127691	0.039799	0.551384
0.1	1	0.134273	0.043627	0.56971	0.131719	0.042692	0.553908
0.2	1	0.133477	0.043096	0.571774	0.129177	0.041106	0.550931
0.3	1	0.134157	0.043373	0.571774	0.129145	0.041125	0.55464
0.4	1	0.134424	0.043731	0.571774	0.128781	0.040995	0.55464
0.5	1	0.136009	0.044221	0.572557	0.128841	0.040975	0.55464
0.6	1	0.136132	0.044262	0.572557	0.127743	0.040477	0.549568
0.7	1	0.136607	0.044352	0.571172	0.128484	0.041047	0.550062
0.8	1	0.136573	0.044345	0.573073	0.128222	0.040971	0.550062
0.9	1	0.136085	0.043953	0.576362	0.128222	0.040947	0.550062
1	1	0.136176	0.043886	0.576362	0.128222	0.040947	0.550062
0	2	0.127571	0.04065	0.558906	0.127571	0.04065	0.558906
0.1	2	0.134279	0.044326	0.567725	0.127598	0.041011	0.548946
0.2	2	0.132227	0.04378	0.565846	0.126087	0.039899	0.548946
0.3	2	0.13283	0.043909	0.57041	0.126327	0.03999	0.55172
0.4	2	0.134523	0.04468	0.57041	0.127085	0.040477	0.55172
0.5	2	0.133141	0.043959	0.57041	0.127483	0.040714	0.55172
0.6	2	0.133539	0.044212	0.57041	0.127142	0.040493	0.55172
0.7	2	0.133565	0.043979	0.57041	0.127077	0.040308	0.55172
0.8	2	0.133548	0.043975	0.571507	0.126337	0.040098	0.55172
0.9	2	0.133639	0.04401	0.571507	0.124933	0.038988	0.54984
1	2	0.13477	0.044337	0.573387	0.124747	0.039042	0.54984

Table 11: Crisis Results-SUnSET with P_{IDF} for *Rel*

		TextRank+*Rel*			TextRank		
beta	EM	AR-1	AR-2	Date-F1	AR-1	AR-2	Date-F1
0	0	0.114358	0.041683	0.354251	0.114358	0.041683	0.354251
0.1	0	0.119213	0.042284	0.364802	0.11779	0.040881	0.363209
0.2	0	0.11937	0.042153	0.365167	0.117337	0.040573	0.363209
0.3	0	0.119812	0.042737	0.369653	0.117364	0.040541	0.363942
0.4	0	0.118798	0.042092	0.370735	0.118033	0.041105	0.367189
0.5	0	0.119503	0.043181	0.375763	0.118011	0.041095	0.367189
0.6	0	0.118468	0.042296	0.37138	0.118011	0.041095	0.367189
0.7	0	0.118511	0.042324	0.370646	0.118011	0.041095	0.367189
0.8	0	0.118789	0.042531	0.369889	0.118011	0.041095	0.367189
0.9	0	0.119232	0.043089	0.373136	0.118011	0.041095	0.367189
1	0	0.118836	0.042856	0.373136	0.118011	0.041095	0.367189
0	1	0.119383	0.044027	0.364554	0.119383	0.044027	0.364554
0.1	1	0.116948	0.041352	0.368236	0.11742	0.041117	0.3668
0.2	1	0.114585	0.040833	0.357094	0.117598	0.041219	0.3668
0.3	1	0.114909	0.041108	0.362224	0.11765	0.041245	0.367534
0.4	1	0.116967	0.041394	0.371358	0.117646	0.041252	0.367534
0.5	1	0.117696	0.042491	0.373769	0.117646	0.041252	0.367534
0.6	1	0.117333	0.042068	0.37128	0.117646	0.041252	0.367534
0.7	1	0.119144	0.043011	0.374527	0.117646	0.041252	0.367534
0.8	1	0.118695	0.042807	0.373769	0.117646	0.041252	0.367534
0.9	1	0.118588	0.042762	0.373769	0.117646	0.041252	0.367534
1	1	0.117075	0.04204	0.372349	0.117646	0.041252	0.367534
0	2	0.114743	0.041232	0.358151	0.114743	0.041232	0.358151
0.1	2	0.113404	0.039178	0.366255	0.112941	0.039047	0.363297
0.2	2	0.112839	0.039618	0.360665	0.112568	0.03918	0.361919
0.3	2	0.112525	0.038941	0.361231	0.113128	0.039125	0.364192
0.4	2	0.112189	0.038723	0.362651	0.113128	0.039093	0.364192
0.5	2	0.113626	0.039906	0.364744	0.113128	0.039093	0.364192
0.6	2	0.112348	0.039172	0.363323	0.112722	0.03876	0.364192
0.7	2	0.112828	0.03947	0.364744	0.11269	0.038761	0.364192
0.8	2	0.112622	0.039416	0.364281	0.11269	0.038761	0.364192
0.9	2	0.112912	0.039573	0.364281	0.11269	0.038761	0.364192
1	2	0.113113	0.039672	0.363477	0.11269	0.038761	0.364192

Table 12: T17 Results-SUnSET with P_{IDF} for *Rel*

		TextRank+*Rel*			TextRank		
beta	EM	AR-1	AR-2	Date-F1	AR-1	AR-2	Date-F1
0	0	0.127151	0.038522	0.551891	0.127151	0.038522	0.551891
0.1	0	0.133361	0.042997	0.571505	0.128408	0.040082	0.55271
0.2	0	0.133766	0.042873	0.575374	0.128416	0.039956	0.554628
0.3	0	0.134439	0.042803	0.575513	0.128083	0.039733	0.556272
0.4	0	0.135246	0.043226	0.575513	0.128149	0.039836	0.556272
0.5	0	0.135214	0.043271	0.574128	0.128149	0.039836	0.556272
0.6	0	0.135432	0.043603	0.573612	0.128149	0.039836	0.556272
0.7	0	0.135341	0.043572	0.575207	0.128149	0.039836	0.556272
0.8	0	0.134963	0.043413	0.575658	0.12811	0.039807	0.556272
0.9	0	0.13518	0.043546	0.574529	0.12811	0.039807	0.556272
1	0	0.133889	0.042507	0.57265	0.12811	0.039807	0.556272
0	1	0.128352	0.040235	0.553264	0.127691	0.039799	0.551384
0.1	1	0.137057	0.044288	0.574131	0.129329	0.040677	0.545189
0.2	1	0.136771	0.043724	0.576053	0.129147	0.040417	0.548139
0.3	1	0.13739	0.043633	0.576029	0.128679	0.040163	0.548139
0.4	1	0.137267	0.043609	0.576029	0.129101	0.040205	0.549784
0.5	1	0.137116	0.043592	0.574128	0.129101	0.040205	0.549784
0.6	1	0.137352	0.043926	0.573612	0.129101	0.040205	0.549784
0.7	1	0.136805	0.043857	0.573612	0.129101	0.040205	0.549784
0.8	1	0.136675	0.043737	0.572515	0.129101	0.040205	0.549784
0.9	1	0.136709	0.043784	0.571386	0.129101	0.040205	0.549784
1	1	0.135355	0.04276	0.569507	0.129101	0.040205	0.549784
0	2	0.127571	0.04065	0.558906	0.127571	0.04065	0.558906
0.1	2	0.134797	0.043389	0.574589	0.126857	0.040582	0.551637
0.2	2	0.134322	0.043568	0.57271	0.125526	0.039409	0.553047
0.3	2	0.134659	0.043783	0.573226	0.125708	0.039558	0.553047
0.4	2	0.134008	0.042938	0.57271	0.125539	0.039507	0.552531
0.5	2	0.13373	0.043024	0.57271	0.12594	0.039747	0.552531
0.6	2	0.13373	0.043039	0.57271	0.12594	0.039747	0.552531
0.7	2	0.134293	0.043185	0.57271	0.12594	0.039747	0.552531
0.8	2	0.134293	0.043185	0.57271	0.12594	0.039747	0.552531
0.9	2	0.134407	0.043212	0.57271	0.12594	0.039747	0.552531
1	2	0.134075	0.043106	0.57271	0.12594	0.039747	0.552531

G Example Timelines Generated from Experiments

Ground Truth

2010-04-20	Deepwater Horizon drilling rig explodes about 42 miles off Louisiana , killing 11 men .
2010-04-22	The rig , having burned and been showered with water during firefighting efforts , sinks .
2010-05-02	The federal government closes 3 percent of federal waters in the gulf to fishing .
2010-05-06	BP tries to stop the spill by lowering a 98-ton `` containment dome '' over the leak .
2010-05-26	BP begins `` top kill '' attempt , shooting mud down the drillpipe in an attempt to clog the leaking well .
2010-05-27	President Obama announces a six-month moratorium on new deepwater drilling in the gulf .
2010-06-02	37 percent of federal waters in the gulf are closed to fishing , the largest extent of the summer 's closures .
2010-07-15	The well is finally closed .
2010-09-17	BP begins `` bottom kill , '' a final injection of cement at the well 's base that rendered it physically incapable of spilling another drop .
2010-10-01	Tony Hayward , BP 's chief executive when the spill began , steps down .
2010-10-12	Obama lifts the moratorium on deep-water drilling ahead of schedule , pledging stricter oversight of safety on rigs .
2010-10-15	The government re-opens nearly 7,000 square miles of federal waters to fishing , the ninth re-opening since the leak was capped .

Figure 8: Example of Ground Truth Timeline of T17 BP-Oil

LLM-TLS

2010-04-29	BP faces a huge clean-up bill and possible damage to its global brand after the Deepwater Horizon rig sank following an explosion and fire.
2010-05-02	Louisiana Governor Bobby Jindal warns that the oil spill threatens the very way of life of people in his state.
2010-05-24	US government officials threaten to keep their boot on BP's neck until the oil spill is plugged.
2010-05-26	BP to make decision on 'top kill' plan to stem oil leak
2010-05-27	President Obama defends his administration's response to the Gulf of Mexico oil spill, but admits that the government does not have the technology to deal with the damaged oil well.
2010-06-01	BP 's stock nose-dived on Tuesday , losing nearly 15 percent of its value on the first trading day since the previous best option -- the so-called top kill -- failed and was aborted at the government 's direction .
2010-06-03	BP oil spill workers complain of symptoms.
2010-06-15	President Obama announces that he will demand that BP \'s chairman set aside whatever money is needed to compensate those who have been harmed by the company \'s "recklessness" in the Gulf of Mexico.
2010-06-16	BP cuts dividend and capex to pay for oil spill
2010-06-17	BP CEO Tony Hayward grilled by Congress over oil spill
2010-07-15	BP says it has temporarily stopped the flow of oil into the Gulf of Mexico from its leaking well.
2010-09-08	BP spreads blame over oil spill

(a) LLM-TLS Generated Timeline

SUnSET

2010-04-20	The Deepwater Horizon rig exploded, killing 11 workers and causing a massive oil spill in the Gulf of Mexico.
2010-04-22	The Deepwater Horizon rig sinks, leading to the start of the Gulf of Mexico oil spill.
2010-04-29	President Obama pledges resources to contain the spill and holds BP responsible.
2010-05-02	Obama visits the Gulf Coast; fishing areas are closed; BP starts drilling a relief well.
2010-05-07	BP's attempt to use a containment dome fails due to frozen hydrocarbons.
2010-05-26	BP initiates the 'top kill' procedure to stop the oil leak by pumping heavy drilling mud into the damaged blowout preventer.
2010-05-27	BP begins the 'top kill' procedure to stop the oil leak, which is expected to take 24 hours or longer. President Obama announces a six-month moratorium on new deepwater drilling permits and orders 33 deepwater exploratory wells to halt operations. The head of the Minerals Management Service resigns amid criticism of lax regulation. New estimates suggest the spill is significantly larger than previously thought, potentially four to five times BP's initial estimate.
2010-06-01	BP launches operation to cut damaged pipe and seal it with a cap to contain the Gulf of Mexico oil spill. US Attorney General Eric Holder announces criminal and civil investigations into the spill. BP's shares fall sharply due to the failure of the 'top kill' method. President Obama describes the leak as the greatest environmental disaster of its kind in our history.
2010-06-15	President Obama delivers his first Oval Office speech addressing the BP oil spill, emphasizing the need for comprehensive cleanup measures and an oil reform program to prevent future disasters and reduce oil use and pollution.
2010-06-16	President Obama meets with BP executives regarding the Gulf oil spill and announces BP\'s agreement to place $20 billion into an escrow fund to cover damages.
2010-06-17	BP CEO Tony Hayward testifies before a House committee on the Gulf oil spill, providing few answers and facing criticism.
2010-07-15	BP successfully places a temporary cap on the leaking well, stopping the flow of oil into the Gulf of Mexico.

(b) SUnSET Generated Timeline

Figure 9: LLM-TLS *(a)* versus SUnSET *(b)* in generating example timeline of T17 BP-Oil (Fig. 8). Text with green highlight indicates event which aligns better with Ground Truth

Ground Truth

2011-03-15	First protests after calls on Facebook for a 'Day of Dignity'.
2011-08-18	US President Barack Obama and his allies urge Assad to quit .
2011-10-02	Creation of the opposition Syrian National Council SNC .
2011-11-16	The Free Syrian Army FSA , set up in Turkey by a deserter from Assad 's army , claims an attack on a Damascus military base .
2012-03-21	The UN Security Council demands that Syria implement a peace plan drawn up by UNArab League envoy Kofi Annan calling for a ceasefire and democratic transition .
2012-04-12	The UNbacked ceasefire , which the regime accepted , formally begins , but violence continues .
2012-05-25	At least 108 people , including 49 children and 34 women , killed in Houla in Homs province .
2012-06-16	UN observer head General Robert Mood suspends the work of his mission because of escalating violence , although the monitors remain in Syria .
2012-07-03	The USbased Human Rights Watch says Syria is holding tens of thousands of detainees in a torture archipelago .
2012-07-06	Munaf Tlass , a highprofile general close to Assad , defects .
2012-07-18	Four top security officials , including Assad 's brotherinlaw Assef Shawkat , are killed in a Damascus attack .
2012-07-19	Russia and China veto for the third time a UN Security Council resolution that would threaten sanctions against Assad .
2012-07-23	Regime forces , boosted by elite troops , reclaim most of the capital after days of unprecedented clashes .
2012-07-28	The military launches a dawn assault on Syria 's economic capital Aleppo , where clashes have been taking place since the 20th , using helicopter gunships , artillery and tanks as well as ground troops .
2012-08-02	Annan resigns after five months of fruitless peace efforts .
2012-08-13	Rebels claim they downed a fighter jet .
2012-08-14	More than 23,000 people have been killed since the outbreak of the revolt , the Syrian Observatory for Human Rights says .
2012-08-15	A bomb explodes in central Damascus near the hotel used by the UN observer mission in Syria , wounding three people .

Figure 10: Example of Ground Truth Timeline of Crisis Syria

LLM-TLS

2011-11-17	Syrian rebels attack a military intelligence complex in Damascus, the first such attack on a government target in the capital.
2011-12-01	Syrian soldiers defect from army as country appears poised for civil war
2011-12-02	The UN human rights chief says Syria risks being engulfed in a civil war unless President Bashar al-Assad's government ends its crackdown on opposition protesters.
2012-01-13	The Arab League suspends Syria's membership and calls for sanctions against the regime.
2012-01-15	Syrian President Bashar Assad issues a general amnesty for crimes committed in the context of the uprising against his regime.
2012-01-31	Russia says a U.N. resolution demanding Syrian President Bashar Assad step aside is a " path to civil war .
2012-02-06	Russia and China veto a UN resolution calling for Syrian President Bashar al-Assad to step down.
2012-02-09	Syrian forces bombard Homs, killing at least 100 civilians
2012-05-08	The United Nations and the Arab League warn that Syria is sliding into civil war as the government and rebels continue to fight despite a ceasefire.
2012-05-28	Syrian President Bashar al-Assad's regime is accused of carrying out a massacre of 108 people in the town of Houla, including 49 children.
2012-05-31	Secretary of State Hillary Clinton warns that Syria is likely to descend into civil war if Russia does not use its influence to push President Bashar al-Assad to step down.
2012-06-03	Syrian President Bashar al-Assad denies that his regime was responsible for the massacre of 108 people in Houla , including 49 children , and blames the killings on " foreign-backed terrorists .
2012-06-12	The United Nations peacekeeping chief says the conflict in Syria has become a civil war.
2012-07-15	The International Committee of the Red Cross declares the conflict in Syria to be a civil war , meaning international humanitarian law applies throughout the country .
2012-07-16	The International Committee of the Red Cross declares that the conflict in Syria is a civil war.
2012-07-18	A suicide bombing in Damascus kills several top Syrian officials, including the defense minister and the president's brother-in-law.
2012-07-27	Syrian troops reportedly strafed several neighbourhoods of Aleppo from helicopter gunships, after an MP defected to Turkey.
2012-08-02	Kofi Annan resigns as the U.N. and Arab League envoy to Syria, citing the failure of the international community to unite over the country's escalating violence

Figure 11: LLM-TLS Generated Timeline of Crisis Syria (Fig. 10)

SUnSET

2011-03-15	Syria's uprising began, marking the start of protests against President Bashar al-Assad's government.
2012-02-04	Russia and China veto a UN resolution aimed at ending the conflict in Syria, backing Assad's regime.
2012-02-22	Journalists Marie Colvin and Remi Ochlik are killed during a rocket attack on a makeshift media center in Baba Amr, Homs.
2012-03-01	Government forces enter Baba Amr, a former opposition stronghold in Homs, after a month-long assault. The area is described as 'completely destroyed' by Valerie Amos, the UN humanitarian chief.
2012-03-15	Thousands rally in Damascus in support of Assad's regime as troops tighten siege in opposition areas on the one-year anniversary of the uprising.
2012-04-12	Annan's ceasefire deal takes effect in Syria, but fighting continues.
2012-05-08	Kofi Annan, international envoy, warns of the risk of civil war in Syria due to increasing violence and bombings. He expresses profound concern from world powers and states that deploying 300 truce monitors by the end of May is crucial to stabilize the situation.
2012-05-25	A massacre in Houla, Syria, results in over 100 deaths, including 49 children under 10, attributed to pro-regime elements and government forces.
2012-05-26	At least 108 people were killed in the Syrian town of Houla, prompting international condemnation including from Kofi Annan who called it an 'appalling crime'.
2012-06-01	Secretary of State Hillary Clinton accuses Russia of contributing to a potential civil war in Syria, and U.S. Ambassador to the UN Susan Rice denounces Russian arms shipments to Syria as 'reprehensible'
2012-06-12	U.N. peacekeeping chief declares Syria's 15-month uprising a full-scale civil war, citing intensified fighting and use of heavy weaponry by Assad's forces
2012-06-13	UN peacekeeping chief declares Syria's conflict a civil war; U.S. expresses concern over Russian arms shipments to Syria; UN observers face hostile reception in Haffeh; at least 51 civilians and 12 soldiers killed across Syria.
2012-06-22	Syria shoots down a Turkish F4 jet over the Mediterranean Sea, leading to heightened tensions between the two countries. Two Turkish pilots are missing. Turkey vows to take necessary steps in response.
2012-07-15	The International Committee of the Red Cross declares the conflict in Syria to be a civil war , meaning international humanitarian law applies throughout the country .
2012-07-18	A bomb attack in Damascus kills four top Syrian security officials, including Assad's brother-in-law Assef Shawkat, Defense Minister Dawoud Rajhah, and the vice president's military adviser Hasan Turkmani.
2012-08-01	4,933 people were killed in Syria during August, marking the deadliest month in the 17-month civil war.
2012-08-02	Kofi Annan resigns as UN peace envoy to Syria, citing the impossibility of his mission due to ongoing violence and lack of international consensus.
2012-08-15	A bomb explodes near a hotel housing U.N. observers in Damascus, injuring at least three people.

Figure 12: SUnSET Generated Timeline of Crisis Syria (Fig. 10)