Enabling Roll-up and Drill-down Operations in News Exploration with Knowledge Graphs for Due Diligence and Risk Management

Sha Wang[1], Yuchen Li[1], Hanhua Xiao[1], Zhifeng Bao[2], Lambert Deng[3] Yanfei Dong[4]
[1] Singapore Management University [2] RMIT University [3] DBS Bank Limited [4] PayPal
sha.wang.2021@phdcs.smu.edu.sg, yuchenli@smu.edu.sg, hhxiao.2020@phdcs.smu.edu.sg, zhifeng.bao@rmit.edu.au, lambertdeng@dbs.com, dyanfei@paypal.com

***Abstract*—Efficient news exploration is crucial in real-world applications, particularly within the financial sector, where numerous control and risk assessment tasks rely on the analysis of public news reports. The current processes in this domain predominantly rely on manual efforts, often involving keyword-based searches and the compilation of extensive keyword lists. In this paper, we introduce NCEXPLORER, a framework designed with OLAP-like operations to enhance the news exploration experience. NCEXPLORER empowers users to use *roll-up* operations for a broader content overview and *drill-down* operations for detailed insights. These operations are achieved through integration with external knowledge graphs (KGs), encompassing both fact-based and ontology-based structures. This integration significantly augments exploration capabilities, offering a more comprehensive and efficient approach to unveiling the underlying structures and nuances embedded in news content. Extensive empirical studies through master-qualified evaluators on Amazon Mechanical Turk demonstrate NCEXPLORER's superiority over existing state-of-the-art news search methodologies across an array of topic domains, using real-world news datasets.**

I. INTRODUCTION

Risk assessment and due diligence are paramount for financial institutions, merchants, and all relevant stakeholders. For instance, global payment giant PayPal conducts Anti-Money Laundering (AML) and Counter-Terrorist Financing (CTF) risk assessments, as well as sanctions screening, for every new business client [1]. Similarly, DBS, Southeast Asia's largest bank, shoulders additional responsibilities in the realm of Environmental, Social, and Governance (ESG) concerns [2]. DBS's ESG risk policy explicitly prohibits financial involvement in activities associated with illegal logging, forced or child labor, wildlife trading, and more. In risk assessments, analysts in financial institutions rely on public news reports to determine whether the entities under scrutiny have ever been associated with suspicious activities. Due to the complexity of these due diligence tasks, they are predominantly manual. Compliance teams laboriously maintain extensive lists of financial crime terminology and sift through search results to distinguish genuine financial misconduct from unrelated or benign news. This creates a big operational overhead of running a business. According to a recent McKinsey survey [3], compliance-related expenses in major banks have surged, becoming unsustainable. The situation becomes even more precarious for DNFBPs (Designated Non-Financial Businesses and Professions), such as precious metal dealers or real estate agents, often operating with limited resources. There is a clear need for a more efficient approach to news exploration.

In our research, we introduce an OLAP-inspired approach to news exploration. In the *roll-up* process, users input known terms, leading to the generation of broader topics. For example, "*FTX*" is expanded to "*Bitcoin Exchange*". Our system, NCEXPLORER, subsequently amplifies these topics by curating a list of relevant keywords for retrieval. The retrieved articles come with an array of related subtopics, granting users the capability to *drill-down* into specific news pieces and discover unanticipated topics.

To illustrate the enhanced productivity gained with our approach, especially in due diligence checks, consider a Know Your Customer (KYC) task involving a newly incorporated cryptocurrency exchange, "*CryptoX*", as depicted in Fig **??**. The exchange seeks to open a business bank account in a jurisdiction with recent digital currency regulations. The KYC analyst begins by querying "*CryptoX fraud*" but finds no results. Realizing "*CryptoX*" has a clean slate, the analyst shifts to peer-related checks using queries such as "*FTX*" and "*Fraud*". This approach yields some results alongside *roll-up* options. Expanding the search to industry-wide topics like "*Bitcoin Exchange*" and "*Financial Crime*" produces a more comprehensive set of results, each linked to entities relevant to the chosen topics (highlighted in color). Armed with this information, the expert delves deeper into understanding the prevalent fraud types in the crypto realm and regulatory implications. Throughout this journey, the analyst enjoys the leeway to alternate between roll-up and drill-down modes, mirroring the flexibility of navigating an OLAP cube. In comparison, traditional due diligence would demand the painstaking manual tweaking of keywords and a thorough examination of search outputs to discern interconnected entities and patterns.

Our proposed OLAP operations for news exploration are versatile and hold potential for many other novel applications. For instance, NCEXPLORER can detect media bias. When Elon Musk acquired Twitter, it stirred debates on wealthy individuals influencing media [4]. By using NCEXPLORER, users can expand from "*Elon Musk*" to unveil parallels like Jeff Bezos's acquisition of the Washington Post [5], Patrick Soon-Shiong's purchase of the Los Angeles Times [6], and Rupert

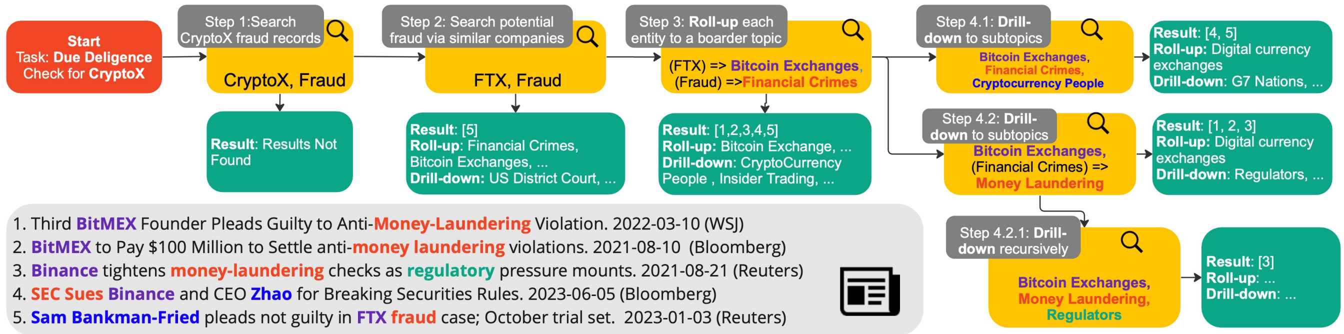

Fig. 1: NCExplorer ***roll-up*** and ***drill-down*** example

Murdoch's buyout of The Wall Street Journal [7]. While articles on Musk leaned negative, others retained a neutral or positive bent. These disparities underscore NCExplorer's prowess in discerning biases and news narratives.

This OLAP-like interaction is made possible through the integration of the ontology network of external Knowledge Graphs (KGs) [8], [9]. As illustrated in Fig 2, the KG encompasses not only millions of entities and relationships, forming a comprehensive fact network, but also extensive ontology information. While vast ontology networks make *roll-up* and *drill-down* operations possible, two major challenges arise. Firstly, in *roll-up* operations, identifying relevant news articles that correspond to selected roll-up concepts is complicated by the fact that roll-up concepts are part of the Knowledge Graph's (KG) ontology space, while the news articles' entities often belong to the KG's instance space. To address this, we've developed a relevance measure linking these two spaces and devised efficient path-based methods to quickly pinpoint relevant articles for roll-up analysis. Secondly, for drill-down operations, the key challenge lies in the effective ranking of sub-topic concepts to facilitate in-depth *drill-down* investigations by analysts. This task is particularly demanding due to the vast array of potentially relevant concepts suitable for drill-down analysis. To meet this challenge, we have designed a ranking framework according to the business objectives, incorporating three critical dimensions: coverage, specificity, and diversity. We summarize our technical contributions as follows:

- We introduce a novel framework, NCExplorer, which is the first of its kind, supporting news exploration with semantic *roll-up* and *drill-down* operations by leveraging the ontology and fact networks of knowledge graphs (KGs).
- We develop effective ranking schemes to evaluate concept-article relevance, enabling smooth *roll-up* and *drill-down* operations. Additionally, we devise an efficient, unbiased sampling estimator for relevance score computation.
- We assessed NCExplorer using extensive news datasets, garnering feedback from Amazon Mechanical Turk's master-qualified participants. With over 3,900 evaluations, the findings affirm NCExplorer's superiority over leading news search techniques. Even after integrating GPT models into all compared baselines, the results remained consistent.
- We also release our implementations, datasets, evaluation results and a full report at [10]. Our dataset contains 200k news articles, with 2.9 million entity and 3.7 million concept annotations linked to DBPedia [11]. Compared to existing news datasets, the inclusion of entity and concept annotations enables deeper analysis of news articles, leveraging the interconnectedness of the knowledge graphs.

II. Related Work

Our work is broadly related to news search and recommendation, aiming to enhance user experience in these domains. Existing works in this area typically extend generic natural language understanding (NLU) models [12]–[14] to improve search and recommendation experiences for news articles. These works generally fall into two categories below:

Embedding-based approaches: They leverage embeddings to create document representations that capture the semantic meaning of news articles. Notable works include DKN [15], KRED [16], and NewsGraph [17]. DKN integrates the embeddings of news titles, derived from a CNN approach [18], with the embeddings of news KG entities obtained through a knowledge graph embedding technique [19]. KRED extends this idea further by introducing a context embedding layer generated from a news entity's neighbors, as well as an attentive merging layer inspired by the Knowledge Graph Attention Network (KGAT). NewsGraph, on the other hand, builds a separate graph by removing irrelevant edges and introducing new edges that represent co-occurrences in the same news or visits from the same user. This alteration hydrates the static KG with dynamic interactions between entities in real life, making the latent vectors aware of current affairs. While embedding-based approaches are shown to improve the semantic representations of documents, they lack an explicit explanation to associate entities mentioned in text with concepts appeared in query, which is crucial for due diligence.

Structure-based approaches: Our work is more related to the methods that focus on utilizing the KG structures to provide better explainability and semantic relevance between news articles. AnchorKG [20] and NewsLink [21] are two prominent works in this category. AnchorKG represents each news article as a compact subgraph containing essential news entities and their k-hop neighbors, generated using a reinforced learning framework. This interaction of anchor graphs provides explanations for semantic relevance between articles.

NewsLink, a state-of-the-art news search approach, uses seed nodes identified from text fragments and a graph expansion algorithm to connect nodes in a single graph, adding hidden related nodes as auxiliary information for news semantic representation. These methods boost search performance via KG's fact network. However, the exploration capability is still limited due to lack of KG ontology integration. To the best of our knowledge, our work is the first to facilitate systematic exploration with OLAP-like operations.

III. NCExplorer

NCExplorer leverages external KGs for news exploration. A KG is a multigraph $\mathcal{G} = (\mathcal{V}_C \bigcup \mathcal{V}_I, \mathcal{E}_C \bigcup \mathcal{E}_I, \Psi)$. $\mathcal{V}_C$ and $\mathcal{V}_I$ represent the concept entities and the instance entities, represented by yellow and green nodes in Fig 2 respectively. Each concept edge in $\mathcal{E}_C$ links two concept entities, and an instance edge in $\mathcal{E}_I$ links two instance entities. Like NewsLink [21], we add a reversed edge for each original edge so that $\mathcal{G}$ is bidirected. The associations between the instance space and the concept space are captured by the ontology relation Ψ. $\Psi(c)$ maps a concept entity $c \in \mathcal{V}_C$ to a set of instance entities while $\Psi^{-1}(v)$ maps an instance entity $v \in \mathcal{V}_I$ to a set of concept entities. Fig. 3 shows the framework's architecture. When a stream of news articles arrives, they undergo a pipeline of natural language processors (NLP) of tokenization, entity recognition and entity linking. We use the spaCy [22] library as our NLP tool to transform a document into a list of KG instance entities. The next step is to expand into the KG ontology space to enable two major components of NCExplorer: semantic *roll-up* and *drill-down*.

A. Roll-up Operation.

NCExplorer generates entities from documents d, allows users to replace some entities with KG concepts, and further *roll-up* along the edge type "*broader*" , to form a concept pattern query Q. Given such a Q, a document d matches Q if for each concept $c \in Q$, there is an entity v in d that $v \in \Psi(c)$. In other words, we can find a set of entities in d that match Q. The relevance score of a matched document d to Q is the sum of the relevance scores among all concepts in Q, i.e.,

$$rel(Q, d) = \sum_{c \in Q} cdr(c, d). \tag{1}$$

Definition 1: [Roll-up] Given a concept pattern query Q, return top-K documents d with the highest score $rel(Q, d)$.

Concept Document Rank. We propose a novel ranking scheme for $cdr(c, d)$ by considering two key relevance dimensions: *ontology relevance* $cdr_o(c, d)$ and *context relevance* $cdr_c(c, d)$. A concept c is relevant to a document d if c is relevant in both dimensions:

$$cdr(c, d) = cdr_o(c, d) \cdot cdr_c(c, d). \tag{2}$$

1) Ontology Relevance: When a document d contains an entity v that matches a concept c under the ontology relation Ψ, i.e., $v \in \Psi(c)$, c is associated with d by ontology relevance. Since there could be more than one entity in d that matches c, we define the ontology relevance score function as follows:

$$cdr_o(c, d) = log\frac{|\mathcal{V}_I|}{|\Psi(c)|} \cdot \left(\max_{v \in ME(c,d)} tw(v, d)\right) \tag{3}$$

where $ME(c, d) = \{v | \; v \in d \text{ and } v \in \Psi(c)\}$. First, a concept that matches more entities in the ontology, i.e., lower specificity, should be less relevant to a document. Second, among the matched entities, a *pivot* entity is selected as the one with the highest term weight $tw(v, d)$ in the document to match the concept. The term weight reflects the importance of v in d. If v plays a more significant role in d, c is more relevant to d. We use the typical TF-IDF scheme for term weighting in implementation and other schemes can be easily supported. For a broad concept that does not have a direct link to document entities, $cdr_o(c, d)$ is replaced with an edge concept among its children that matches document entities .

2) Context Relevance: An entity in a document could map to different concepts by ontology. Hence, ontology relevance can only differentiate the matched concepts of the same entity with the specificity score $\log(|\mathcal{V}_I|/|\Psi(c)|)$, which simply penalizes broad concepts. We thus propose to include the unmatched entities as contextual information to improve the relevance semantics. To measure the context relevance of c and d, we compute the relevance of c and the context entities $CE(c, d) = \{v | \; v \in d \text{ and } v \notin \Psi(c)\}$. Although the context entities do not match c, we take advantage of both the ontology relation and the instance space to link c with the context entities. We introduce a novel connectivity score $conn(c, d)$ to measure the KG connectivity between a concept c and a context entity set $CE(c, d)$ as follows:

$$conn(c, d) = \sum_{v \in CE(c,d)} \frac{\sum_{u \in \Psi(c)} \sum_{l=1}^{\tau} \beta^l \cdot |paths_{u,v}^{<l>}|}{|CE(c, d)|} \tag{4}$$

where $|paths_{u,v}^{<l>}|$ measures the number of l-hop simple paths connecting u and v in the instance space, and β is a damping factor that penalizes longer paths. Intuitively, the connectivity score is the average number of paths among all context entities connected to any KG instance entity that matches c, subject to a hop constraint of at most τ. Thus, better connectivity leads to a higher relevance score. Finally, we normalize the connectivity score to $[0, 1)$ as follows:

$$cdr_c(c, d) = 1 - \frac{1}{1 + conn(c, d)} \tag{5}$$

B. Drill-down Operation.

Based on the matched news from the *roll-up* operation, NCExplorer suggests additional concepts as subtopics to a query Q. A suggested concept c', as a subtopic of Q, enables users to conduct a *drill-down* analysis. By selecting c', users narrow down the matched news to the augmented query $Q \cup c'$. Since a subtopic must appear as a concept in at least one

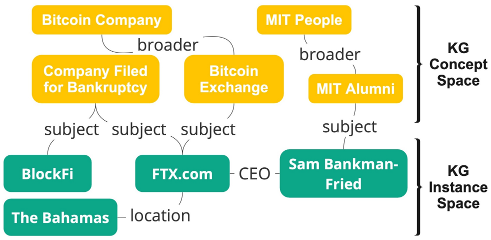

Fig. 2: KG Concept and Instance Spaces

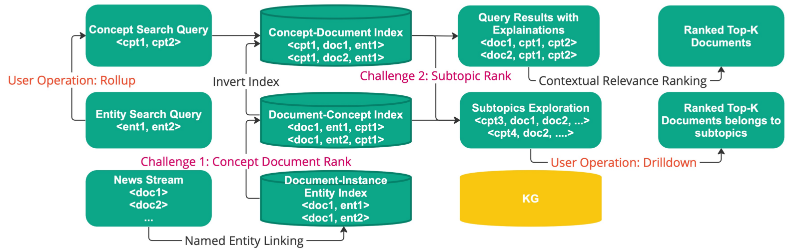

Fig. 3: NCEXPLORER Architecture.

matched document of Q, we can find all candidate subtopics by unionizing the concepts from all matched documents.

Let $\mathcal{D}(Q)$ denote the set of retrieved documents for Q and a candidate subtopic is a concept c such that there is an instance entity $v \in \Psi(c)$ appearing in one of the documents $d \in \mathcal{D}(Q)$. We develop a score $sbr(\cdot, Q)$ over candidate subtopics and only suggest top-scored subtopics for Q. There are three key perspectives considered in $sbr(\cdot, Q)$ as below:

$$sbr(c, Q) = coverage(c, Q) \cdot specificity(c) \cdot diversity(c, Q)$$

1) Coverage: The coverage score of a subtopic c is calculated as the sum of the relevance scores $cdr(c, d)$ for all documents d in $\mathcal{D}(Q)$. This ensures that only subtopics that are highly relevant to a large number of documents are suggested to the user for further exploration: $coverage(c, Q) = \sum_{d \in \mathcal{D}(Q)} cdr(c, d)$.

2) Specificity: To avoid suggesting trivial subtopics like "*Person*" which may match a large number of entities, we prioritize concepts with higher specificity scores using the formula $specificity(c) = \log(|\mathcal{V}_I|/|\Psi(c)|)$.

3) Diversity: The diversity score calculates the average number of distinct entities mapped to the concept c among all the documents relevant to $Q \cup \{c\}$: $diversity(c, Q) = \frac{|\bigcup_{d \in \mathcal{D}(Q)} ME(c,d)|}{|\mathcal{D}(Q \cup \{c\})|}$. This helps ensure fairness in the suggested concepts and prevents results from being biased toward concepts that match a small set of popular entities.

Definition 2: [Drill-down] Given a concept pattern query Q, return top-K subtopics c with the highest $sbr(c, Q)$ score.

C. Connectivity Score Estimation

The connectivity score can be computationally expensive due to the massive number of path enumerations for each pair of entities and existing studies on s-t path enumeration can only produce *polynomial delay* algorithms [23], [24]. Inspired by the sampling approaches for online join aggregation queries [25]–[27], we devise a single random walk estimator. To start a random walk, we first sample a source entity u from all entities that map to the concept c, and a target entity v from the corresponding context entities. A *non-repeating* random walk $r(u, v)$ from u tries to reach v samples via at most τ distinct nodes. Let u_i denote the ith node sampled by a random walk starting from the source u and $N(u_i)$ is the number of eligible neighbors of u_i to be sampled, we define the following sample estimator:

$$r(u, v) = \mathcal{I}(u, v) \cdot |\Psi(c)| \cdot \beta^{l-1} \cdot \prod_{i=1}^{l-1} N(u_i) \tag{6}$$

Algorithm 1: Random walk estimator with k-hop index.

```
Input : KG 𝒢, concept c, document d, sample size θ
Output: the estimated connectivity est
est ← 0;
while there are less than θ random walks do
    l ← 1; p ← 1.0;
    u_l ← a random entity in Ψ(c);
    v ← a random entity in CE(c, d);
    while l ≤ τ do
        count ← 0;
        foreach n ∈ N(u_l) do
            if hop(n, v) ≤ τ − l then
                count ← count + 1;
                u_{l+1} ← n with probability 1/count;
        p ← p/count;
        l ← l + 1;
        break on u_l = v;
    if u_l = v then
        est ← est + β^{l−1}/p;
return est·|Ψ(c)|/θ.
```

where $\mathcal{I}(u, v)$ is an indicator random walk: it returns 1 if u reaches v at the l-th sampled node for any $l \leq \tau$; otherwise, it returns 0. The sampling process may suffer from slow convergence, especially when many of the sampled paths cannot reach the target context entity v. To enable faster convergence, we build a reachability index [28] on the KG instance space and only sample eligible neighbors that satisfy the hop constraint. The following theorem ensures our sample approach is unbiased.

Theorem 1: $r(\cdot, \cdot)$ is an unbiased estimator to $conn(c, d)$, i.e., $\mathbb{E}[r(\cdot, \cdot)] = conn(c, d)$ where the randomness is taken over the source u and target v as well as the random walk from u to v.

Proof Sketch. $conn(c, d)$ is a weighted sum of all the paths connecting an entity node $u \in \Psi(c)$ and a context entity $v \in CE(c, d)$ subject to a hop constraint of τ. For each path $s = \{u_1 = u, u_2, u_3, ..., u_l = v\}$ that connects u and v, the probability of s being sampled is $\mathbb{P}(s) = (|\Psi(c)| \cdot |CE(c, d)| \cdot \prod_i^{l-1} N(s_i))^{-1}$. Thus, we can use the Horvitz–Thompson estimator [29] to approximate the population sum without bias.

Convergence Optimization. Although we devise an unbiased estimator for the connectivity score, the sampling process may suffer from slow convergence, especially when many of the sampled paths cannot reach the target context entity v. To enable faster convergence, we build a reachability index [28] on the KG instance space and only sample eligible neighbors that satisfy the hop constraint. We present our sampling approach with the k-hop reachability index in Algorithm 1. For each random walk, we first sample the source u_1 and the target v. Subsequently, we iteratively sample the next node u_{l+1} by scanning the neighbors of u_l. With the help of the k-hop index, we first check if a neighbor n can reach v within the hop constraint of $\tau - l$ (Line 9) Then, we uniformly sample eligible neighbors by a Reservoir sampler, which only requires a single scan of the neighbors (Line 11) A random walk terminates when either the target v is sampled or the hop limit is reached.

IV. Evaluation

We present the setup in Sec. IV-A and then answer the following questions:

1) Can NCExplorer produce relevant results given topics rolled up from news articles? (Sec. IV-B1)
2) Can NCExplorer efficiently process large-scale news corpus and knowledge graphs? (Sec. IV-B2)
3) Can *context relevance score* effectively measure the relevance between a concept and a document? How does the proposed sampling method impact the accuracy? (Sec. IV-B3)
4) How effective is each scoring component in the ranking model for the *drill-down* operation? (Sec. IV-C)
5) How the combination of *roll-up* and *drill-down* operations improve the productivity of due diligence and risk management tasks in financial institutions.(Sec.IV-D)
6) Can NCExplorer support real analytical applications? (Sec. IV-F)

A. Settings

Datasets. We use the June 2021 snapshot of DBPedia [11] as our backend KG. We crawl 200k articles from popular news portals: *Reuters* [30], *SeekingAlpha* [31] and *The New York Times* [32] to have a mixture of business and politics reports. Detailed statistics of released dataset are shown below.

News Source	Articles	Total Entities	Linked Entities
Seekingalpha	6823	97k	62k (63.9%)
NYT	3625	51k	35k (68.6%)
Reuters	171662	4539k	2336k (51%)

Compared Methods.

- Lucene implements a typical bag-of-words keyword match model. We use BM25 [33] for the term weighting scheme with the default library settings.
- Bert [12] is a popular neural text embedding model. We use SBERT [34], a modification of the pre-trained Bert to map each news article to a vector of 768 dimensions.
- NewsLink [21] is the state-of-the-art implicit news exploration method that expands a news document and a query by forming a common ancestor graph extracted from the KG. Each KG entity in the extracted graph is then treated as a matching keyword in the bag-of-words model.
- NewsLink-Bert is a hybrid method that combines NewsLink and Bert. It expands query entities into a subgraph using NewsLink's algorithm and concatenates them to form a long text query.
- NCExplorer is our proposed approach. The parameters are set to $\tau = 2$ and $\beta = 0.5$ by default. The number of samples for connectivity score estimation is set to 50.

Implementation. NCExplorer and NewsLink are implemented in Python 3.9. Bert and NewsLink-Bert use Qdrant [35] as vector search engine. We use a server running on Ubuntu 20.04 with an AMD EPYC 7643 Processor @ 3.45GHz and 251G RAM.

B. Roll-up Operation

1) Concept Document Rank Evaluation: As shown in Table I, we evaluate a total of six topics. Each topic is combined with either an entity group (can be a list of countries or companies) to form queries such as "*Elections in African countries*" or "*Lawsuits involving U.S. technology companies*". For each query, the top 5 news articles are retrieved from each method, resulting in 25 outcomes. These results are presented to each evaluator in a randomized order. To ensure a fair comparison, we hide NCExplorer's result explanations. The relevance level is rated for each concept in the query, with values ranging between 0 and 5. In total, we obtain 3,900 ratings from 78 evaluators. We use **NDCG@K** to evaluate the effectiveness of relevance ranking. The GPT models [36], [37] present unprecedented capabilities for textual analysis. To explore whether we can use GPT to further boost the performance, we feed the top-5 results from each method into GPT 3.5-turbo [36], ranking the relevance between a topic and a news article based on the prompt:

<news article>. Is this article related to <topic>. please give a rating between 0.000 and 5.000, with 5.000 being most relevant. only give three decimal digits.

Each method's top-K is sorted using GPT's rating and the new **NDCG@K** scores are displayed on the right-hand side. The evaluation result for both scenarios shows that NCExplorer achieves the best or second-best performance in nearly all cases, except for the query "*Merger & Acquisition, U.S. biotechnology companies*". A detailed analysis reveals that evaluators show greater confidence in commonly known surface words like "*M&A*", "*acquisition*", "*buy*", and "*sell*", while expressing uncertainty about specialized terms such as "*takeover*". This finding highlights the effectiveness of NCExplorer's roll-up operation in helping analysts collect news articles featuring more domain-specific vocabularies. Embedding models also demonstrate remarkable performance in news exploration tasks, particularly when combined with KG information. However, there are two issues that arise when using pure embedding models. First, implicit matching may retrieve reports like daily trade price/volume [38], which, although crucial for real-time decision-making, have limited

TABLE I: NCDG@K without/with the GPT model. The best results are boldfaced and the second-best results are underlined.

	NDCG@1 wo/w the GPT model	NDCG@5 wo/w the GPT model	NDCG@10 wo/w the GPT model	NDCG@1 wo/w the GPT model	NDCG@5 wo/w the GPT model	NDCG@10 wo/w the GPT model
Topic	**International Trade**			**Lawsuits**		
Lucene	0.688 / 0.572	0.557 / 0.532	0.737 / 0.720	0.574 / 0.571	0.593 / 0.597	0.763 / 0.766
BERT	0.856 / 0.856	0.882 / 0.882	0.951 / 0.951	**0.849** / 0.849	0.848 / 0.900	0.935 / 0.949
NewsLink	0.765 / 0.817	0.623 / 0.650	0.781 / 0.799	0.329 / 0.628	0.406 / 0.475	0.636 / 0.683
NewsLink-BERT	0.825 / 0.836	0.877 / 0.878	0.949 / 0.949	0.627 / 0.658	0.796 / 0.848	0.883 / 0.898
NCEXPLORER	**0.974 / 0.974**	**0.957 / 0.956**	**0.987 / 0.986**	0.844 / **1.000**	**0.919 / 0.948**	**0.959 / 0.979**
Topic	**Elections**			**Mergers & Acquisitions**		
Lucene	0.550 / 0.273	0.455 / 0.378	0.653 / 0.603	0.464 / 0.659	0.593 / 0.629	0.777 / 0.802
BERT	0.887 / 0.910	0.894 / 0.903	0.941 / 0.947	**0.728 / 1.000**	0.820 / **0.879**	0.915 / 0.955
NewsLink	0.554 / 0.554	0.450 / 0.466	0.649 / 0.660	0.305 / 0.493	0.445 / 0.471	0.675 / 0.693
NewsLink-BERT	**0.946** / 0.946	**0.972 / 0.972**	**0.990 / 0.990**	0.724 / 0.912	0.803 / 0.829	0.913 / 0.930
NCEXPLORER	0.924 / **0.947**	0.958 / 0.966	0.978 / 0.984	0.712 / 0.846	**0.843** / 0.870	**0.937 / 0.956**
Topic	**International Relations**			**Labor Dispute**		
Lucene	0.896 / 0.650	0.722 / 0.670	0.830 / 0.795	0.564 / 0.621	0.618 / 0.632	0.817 / 0.826
BERT	0.921 / 0.957	0.922 / 0.941	0.959 / 0.970	0.370 / 0.370	0.411 / 0.410	0.670 / 0.669
NewsLink	0.735 / 0.804	0.729 / 0.760	0.834 / 0.854	0.481 / 0.729	0.476 / 0.499	0.716 / 0.733
NewsLink-BERT	0.867 / 0.945	0.943 / 0.954	0.971 / 0.982	0.695 / 0.905	0.720 / 0.767	0.889 / 0.922
NCEXPLORER	**0.927 / 0.963**	**0.970 / 0.974**	**0.986 / 0.989**	**0.922 / 0.931**	**0.984 / 0.988**	**0.989 / 0.991**

TABLE II: Impact of the GPT mode

	NDCG@1	NDCG@5	NDCG@10
Lucene	-10.44%	-2.77%	-1.40%
BERT	7.18%	2.89%	1.32%
NewsLink	27.01%	6.20%	3.08%
NewsLink-BERT	7.05%	2.17%	1.05%
NCEXPLORER	6.75%	1.26%	0.84%

TABLE III: *roll-up* and *drill-down* effectiveness study.

Task No.	Keyword Search (avg/std)	NC-explorer (avg/std)	p-value of H1 (n=10)
1	0.6/1.07	2.7/1.56	0.016
2	0.5/0.97	4.0/1.41	0.004
3	0.9/0.99	2.8/0.46	0.001
4	0.9/1.10	2.7/0.67	0.001
5	0.7/0.95	1.8/0.42	0.007
6	1.0/1.56	4.5/1.65	0.002
7	1.0/1.41	4.2/1.03	0.001
8	1.8/1.03	3.3/0.95	0.002

utility in exploration tasks. Second, the issue emerges when the concept entity isn't frequently mentioned alongside instance entities. For example, "*labor dispute*" is often reported in the news as strikes organized by various labor unions. The embedding of "*labor dispute*" alone cannot yield high-quality results. This issue can be partially mitigated with supplementary context provided by Newslink's subgraph KG embedding. For NEWSLINK, the performance is not stable due to the infrequent formation of densely connected single components by the subgraph embedding of multiple concept entities. Instead of identifying hidden nodes that connect existing query entities, the subgraph often results in a single concept entity accompanied by its N-hop neighbors. This dilutes the significance of concepts with smaller connected components, affecting the overall effectiveness of the news exploration process. Table II shows the impact of incorporating the GPT model to re-rank retrieved documents. One observation is that the impact is positive for all methods except LUCENE. Another interesting observation is the impact for NDCG@1 $>$ NDCG@5 $>$ NDCG@10, suggesting that GPT can distinguish the subtle differences among top results. In this work, we only use GPT to re-rank final results instead of performing relevance ranking on all documents. Whether it is feasible to use GPT directly as a relevance ranker is a topic for our upcoming research.

Rating Scale Impact Study Existing work has shown that the effect of scale used in relevance ranking study for crowd-sourced participants do not affect final results [39]. To validate this finding for GPT reranker. We repeated above GPT experiment with 0-10 scale. An independent samples t-test against original 0-5 scale experiment shows a p-value of 0.878. Our chosen significance level (alpha) is 0.05. The p-value is bigger than alpha, rejecting the null hypothesis, showing different scales do not affect evaluation results.

2) Efficiency Study: **Indexing Efficiency.** NCEXPLORER processes each news document and constructs an index for query processing. To evaluate the indexing overhead, we select 100 articles from each news portal and report the average processing time in Fig. 4. LUCENE and BERT show sub-second execution time. NEWSLINK, NEWSLINK-BERT, and NCEXPLORER cost 2-3 seconds for each article. A breakdown analysis of the total cost showed that the top two overheads come from entity linking (91.8%) and relevance score calculation (7.1%). Entity linking is a common cost to all methods that require KG analysis. The primary focus of this work is not to improve the efficiency of entity linking. In fact, the calculation of relevance scores, which constitutes 7.1% of the total indexing time, benefits from our efficient estimation approach using sampling techniques. For constructing the reachability index on the DBpedia KG, comprising 5.2 million nodes and 27.9 million edges, the process takes 260 seconds and necessitates 100GB of memory.

Retrieval Efficiency. Fig 7 displays the query efficiency of the compared methods when increasing the number of concepts while keeping corpus size fixed. The processing time for each data point is the average across 100 queries. LUCENE is a highly optimized system for text retrieval and takes the least amount of time. BERT used to take longer execution time due to the absence of index. Recent development on vector databases has greatly sped up embedding retrieval and result in LUCENE compatible speed. The performance for NCEXPLORER is similar to NEWSLINK where the duration correlates to number of KG entities in the query. NEWSLINK-BERT takes the sum of BERT and NEWSLINK. Overall, NCEXPLORER can answer a concept pattern query with reasonable overhead. Fig. 8 shows the relationship between retrieval time and corpus size for the set of queries. LUCENE and BERT achieve the best scalability thanks to open source solutions. NCEXPLORER and NEWSLINK have higher overhead because the queries are performed on a document database instead of search engines. It is possible improve KG based methods by using LUCENE to retrieve the matched documents before ranking the results with database records. The implementation is skipped in this study as the sub-second retrieval time does not degrades system usability.

3) Context Relevance Score: **Effectiveness and Parameter Study**. To verify the effectiveness of the context relevance

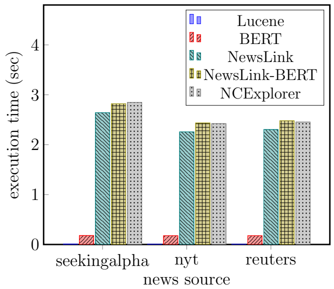

Fig. 4: Indexing time

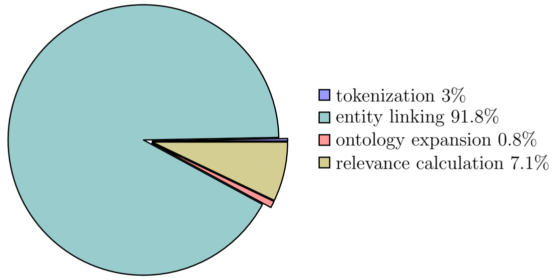

Fig. 5: Document processing time

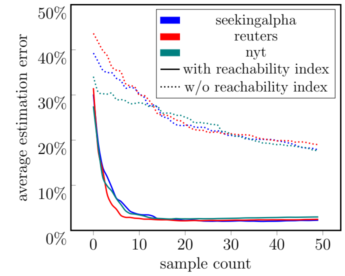

Fig. 6: Sampling error study

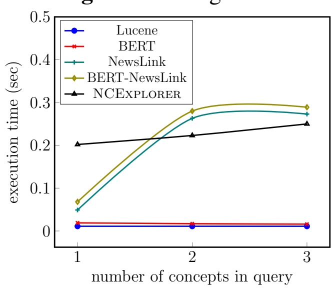

Fig. 7: Retrieval time

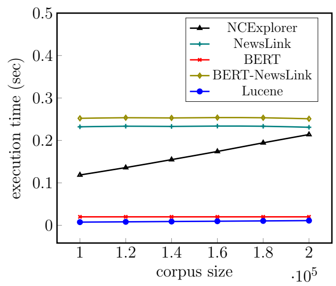

Fig. 8: Scalability Study

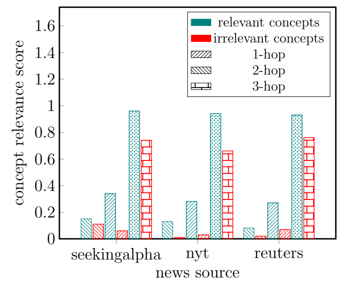

Fig. 9: Context relevance study

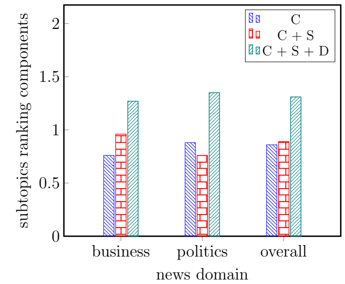

Fig. 10: Subtopics ablation study

score $cdr_c(c, d)$ (Eqn. 4), we design a "negative sampling" approach. We randomly select 100 entries from the inverted index $\langle c, d\rangle$. For each entry $\langle c, d\rangle$, we sample a concept node c' from the KG to generate a "*negative*" concept. Fig. 9 demonstrates the effective differentiation between c and c' using the context relevance score, i.e., $cdc_c(c, d) > cdc_c(c', d)$, regardless of the hop constraint τ from 1 to 3. Notably, when τ is set to 1 and 2, the score differences are significant compared with $\tau = 3$, suggesting that a large τ may include irrelevant concepts due to a higher chance of linking concepts with documents. Our default τ is set to 2, as over half (55%) of the relevant scores are 0 when $\tau = 1$. In contrast, only 22.4% of the relevance scores are 0 for $\tau = 2$, striking a good balance between information linking and relevance differentiation.

C. Drill-down Operation

Ablation Study. Given a search query, NCExplorer automatically suggests related concepts for *drill-down* operations. NCExplorer ranks the concepts by considering three key factors: *Coverage (C)*, *Specificity (S)* and *Diversity (D)*. To investigate the impact of each factor, we conduct user studies for the ablation analysis. We use the same queries from Sec. IV-B1 and select top-ranked concepts when only considering: (1) C; (2) C+S; (3) C+S+D. We build an interactive survey interface (listed in full report) that allows the participants to click on different concepts and view associated results before assigning a distinct rating 1-3 for each subtopic. We recruited participants from the same crowd-source platform AMT [40] and obtain 518 survey results in total. The results are displayed in Fig. 10. We can observe that the specificity contributes slight improvement to the overall rating while diversity plays a more significant role in rating improvement.

D. Roll-up & Drill-down effectiveness study

To assess the effectiveness of *roll-up* and *drill-down* in terms of productivity gain at work, we worked with our compliance team to create a task list resembling a standard set of investigative inquiries. The participants are required to investigate issues like: *Find out the **names of Switzerland Banks** with reports related to **money laundering***. Given the open-ended nature of these tasks, we set a fixed duration of 2 minutes and use the number of accurate responses as a performance metric. This task design mirrors the requirement of Suspicious Activity Reports (SARs) [41], where financial institutions must file findings within limited time of identifying potential criminal activity. The list of questions are:

1) Find out the **categories of commercial crime** involves **Credit Suisse**.
2) Find out the **categories of commercial crimes** involves **Banks of Switzerland**.
3) Find out the **names of Switzerland Banks** with reports related to **commercial crimes**.
4) Find out the **names of Switzerland Banks** with reports related to **money laundering**.
5) Find out the **categories of commercial crime** involves **FTX**.
6) Find out the **categories of commercial crimes** involves **Bitcoin exchanges**.
7) Find out **the names of Bitcoin exchanges** with reports related to **commercial crimes**.
8) Find out the **names of Bitcoin exchanges** with reports related to **money laundering**.

We recruited 10 financial professionals for this study. Using NCExplorer, professionals can generate more answers within the same time limit than the existing keyword match-based method adopted in the corporation, as shown in Table III. We also asked a subjective question "*How likely are you to recommend this tool to due diligence professionals?*" and got an average rating of 8.1/10.

E. Connectivity Score Estimation

RW Estimator Convergence. We evaluate the convergence rate of the proposed RW estimator on the connectivity score. Fig. 6 shows the average sampling error of $cdr_c(c, d)$ compared to the ground truth value. Solid and dotted lines represent RW *with* and *without* the guidance of k-hop reachability index respectively. With the k-hop index, our sampling approach can converge on all three datasets within 5% estimation error using 20 sampling iterations.

F. Case Study on Media Bias

In this case study, NCExplorer's capabilities in identifying media bias are demonstrated. When Elon Musk announced his interest in acquiring Twitter, there were intense criticisms towards rich people controlling the media [4] With NCExplorer, users can roll-up "*Elon Musk*" and "*Twitter*" to a concept pattern query "*American billionaire*" and "*U.S. Mass Media Company*". Among all matched news of the query in the dataset, several similar acquisitions are discovered, such as Jeff Bezos acquiring the Washington Post, Patrick Soon-Shiong buying the Los Angeles Times, and Rupert Murdoch purchasing The Wall Street Journal. Furthermore, the sentiment score of each matched news article is evaluated using a pre-trained model [42]. Interestingly, Musk's acquisition is the only news with a negative sentiment. The rest have either neutral or positive sentiments (Fig 12). Such insights reveal a potential news bias that deserves further investigation by experts. This example showcases how NCExplorer can be a powerful tool for uncovering hidden patterns and biases in news coverage.

V. Conclusion

NCExplorer is a useful tool designed to enhance the news article exploration experience by using OLAP-like operations to connect them with relevant concepts in a knowledge graph. Its ranking system considers both concept relevance and article context, aiding in a comprehensive understanding of the information. NCExplorer not only streamlines due diligence tasks but also improves a range of news analytics tasks. Its efficacy has been validated through crowd-sourced evaluations with a dataset of real-world news articles and a large knowledge graph. Its modular design ensures compatibility with various text-based systems, such as search engines and literature databases, facilitating information discovery and comprehension. Additionally, the release of 200,000 news articles, along with their KG annotations and concept relevance scores, allows for easy customization of the tool to meet specific research needs.

References

[1] Paypal: Anti-money laundering and know your customer. Accessed: 2023-10-01. [Online]. Available: https://publicpolicy.paypal-corp.com/issues/anti-money-laundering-know-your-customer

[2] (2021, Apr.) Dbs bank: Our approach to responsible financing. Accessed: 2023-10-01. [Online]. Available: https://www.dbs.com/iwov-resources/images/sustainability/responsible-banking/DBS%20Bank_Our%20Approach%20to%20Responsible%20Financing_Updated_26Apr2021.pdf

[3] (2019) The compliance function at an inflection point. [Online]. Available: https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/the-compliance-function-at-an-inflection-point

[4] G. Bensinger. (2022) Twitter under elon musk will be a scary place. Accessed: 2023-05-22. [Online]. Available: https://www.nytimes.com/2022/04/25/opinion/editorials/twitter-elon-musk.html

[5] E. Bell. (2013) Jeff bezos' shocking washington post buy was not a business deal — it was a cultural statement. Accessed: 2023-05-02. [Online]. Available: https://twitter.com/BusinessInsider/status/364551288845254656

[6] J. R. K. Meg James. (2018) Billionaire patrick soon-shiong reaches deal to buy l.a. times and san diego union-tribune. Accessed: 2023-05-02. [Online]. Available: https://www.latimes.com/business/hollywood/la-fi-ct-los-angeles-times-sold-20180207-story.html

[7] M. K. Sarah Ellison. (2007) Murdoch wins his bid for dow jones. Accessed: 2023-05-02. [Online]. Available: https://www.wsj.com/articles/SB118589043953483378

[8] T. P. Tanon, G. Weikum, and F. Suchanek, "Yago 4: A reason-able knowledge base," in *ESWC*, 2020, pp. 583–596.

[9] e. a. Lehmann, "Dbpedia, a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic web*, 2015.

[10] Ncexplorer source code. [Online]. Available: https://github.com/knowledge-fusion/ncexplorer/

[11] Dbpedia snapshot 2021-06 release. Accessed: 2023-02-22. [Online]. Available: https://www.dbpedia.org/blog/snapshot-2021-06-release/

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[14] H.-S. Sheu and S. Li, "Context-aware graph embedding for session-based news recommendation," in *RecSys*, 2020, pp. 657–662.

[15] H. Wang, F. Zhang, X. Xie, and M. Guo, "Dkn: Deep knowledge-aware network for news recommendation," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 1835–1844.

[16] D. Liu, J. Lian, S. Wang, Y. Qiao, J.-H. Chen, G. Sun, and X. Xie, "Kred: Knowledge-aware document representation for news recommendations," in *Proceedings of the 14th ACM Conference on Recommender Systems*, 2020, pp. 200–209.

[17] D. Liu, T. Bai, J. Lian, X. Zhao, G. Sun, J.-R. Wen, and X. Xie, "News graph: An enhanced knowledge graph for news recommendation." in *KaRS@ CIKM*, 2019, pp. 1–7.

[18] J. Kim, A.-D. Nguyen, and S. Lee, "Deep cnn-based blind image quality predictor," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 1, pp. 11–24, 2018.

[19] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "Knowledge graph embedding via dynamic mapping matrix," in *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, 2015, pp. 687–696.

[20] D. Liu, J. Lian, Z. Liu, X. Wang, G. Sun, and X. Xie, "Reinforced anchor knowledge graph generation for news recommendation reasoning," in *KDD*. ACM, 2021, pp. 1055–1065.

[21] Y. Yang, Y. Li, and A. K. Tung, "Newslink: Empowering intuitive news search with knowledge graphs," in *ICDE*, 2021, pp. 876–887.

[22] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, vol. 7, no. 1, pp. 411–420, 2017.

[23] L. Qin, Y. Peng, Y. Zhang, X. Lin, W. Zhang, and J. Zhou, "Towards bridging theory and practice: hop-constrained st simple path enumeration," in *International Conference on Very Large Data Bases*. VLDB Endowment, 2019.

[24] S. Sun, Y. Chen, B. He, and B. Hooi, "Pathenum: Towards real-time hop-constrained st path enumeration," in *SIGMOD*, 2021, pp. 1758–1770.

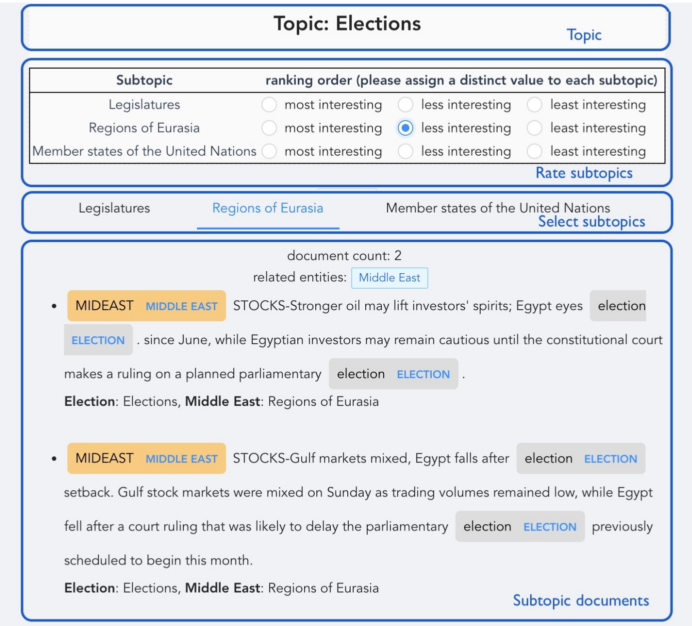

Fig. 11: Subtopic survey interface

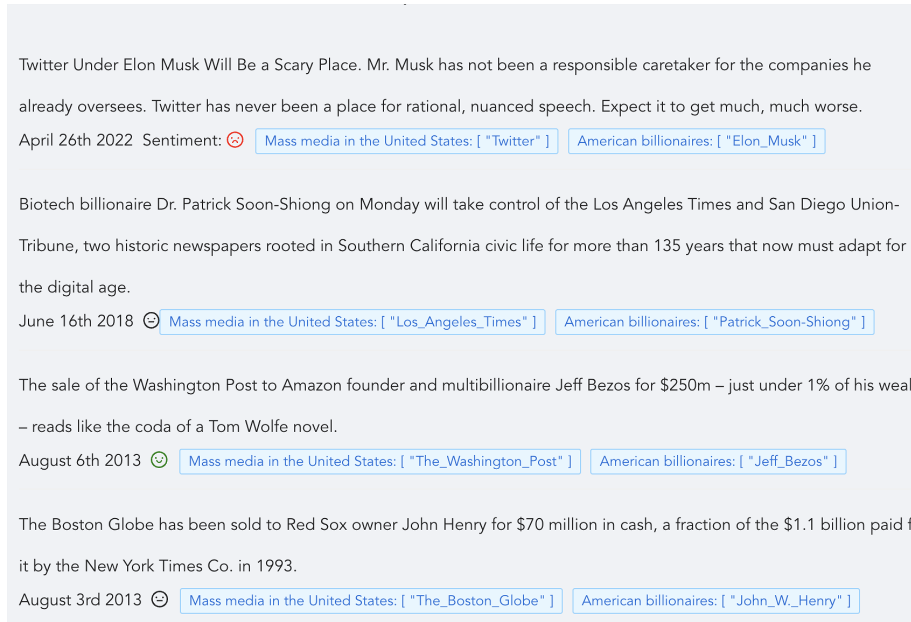

Fig. 12: Case study on Media Bias

[25] F. Li, B. Wu, K. Yi, and Z. Zhao, "Wander join: Online aggregation via random walks," in *SIGMOD*, 2016, pp. 615–629.

[26] Z. Zhao, R. Christensen, F. Li, X. Hu, and K. Yi, "Random sampling over joins revisited," in *SIGMOD*, 2018, pp. 1525–1539.

[27] Y. Park, S. Ko, S. S. Bhowmick, K. Kim, K. Hong, and W.-S. Han, "G-care: a framework for performance benchmarking of cardinality estimation techniques for subgraph matching," in *SIGMOD*, 2020, pp. 1099–1114.

[28] J. Cheng, Z. Shang, H. Cheng, H. Wang, and J. X. Yu, "Efficient processing of k-hop reachability queries," *The VLDB journal*, vol. 23, no. 2, pp. 227–252, 2014.

[29] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, vol. 47, no. 260, p. 663–685, 1952.

[30] Reuters. Accessed: 2023-02-22. [Online]. Available: https://www.reuters.com

[31] Seeking alpha. Accessed: 2023-02-22. [Online]. Available: https://seekingalpha.com

[32] The new york times. Accessed: 2023-02-22. [Online]. Available: https://www.nytimes.com

[33] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[34] Sentence transformers all-mpnet-base-v2. [Online]. Available: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[35] "Vector search engine." [Online]. Available: https://qdrant.tech/

[36] Openai models gpt-3.5. [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5

[37] Openai models gpt-4. [Online]. Available: https://platform.openai.com/docs/models/gpt-4

[38] Bloomberg. Market futures. Accessed: 2023-05-02. [Online]. Available: https://www.bloomberg.com/markets/stocks/futures

[39] K. Roitero, E. Maddalena, S. Mizzaro, and F. Scholer, "On the effect of relevance scales in crowdsourcing relevance assessments for information retrieval evaluation," *Information Processing Management*, vol. 58, no. 6, p. 102688, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457321001734

[40] I. Amazon Mechanical Turk. Amazon mechanical turk. [Online]. Available: https://www.mturk.com

[41] T. O. of the Comptroller of the Currency. Suspicious activity reports (sar). Accessed: 2024-02-02. [Online]. Available: https://www.occ.treas.gov/topics/supervision-and-examination/bank-operations/financial-crime/suspicious-activity-reports/index-suspicious-activity-reports.html

[42] J. M. Pérez, J. C. Giudici, and F. Luque, "pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks," 2021.