Trustworthy Source Extraction via Topic-Related Named Entity Recognition and Higher-Order Singular Value Decomposition

Anonymous ACL submission

Abstract

Understanding the trustworthiness of sources is paramount in the domains of text retrieval, generation, and numerous other tasks, particularly in the context of social media where unverified information can have detrimental effects. Solely relying on semantic analysis proves insufficient for assessing source credibility, underscoring the necessity for advanced information extraction techniques and cross-validation across diverse sources. In this study, we present an elegant yet efficient model to analyze the intricate network of source relationships and key information attributes. Utilizing advanced named entity recognition specialized in key information extraction, we methodically distill critical data to assemble a multidimensional tensor. This tensor comprehensively represents the intricate web of interactions among sources and the essential informational elements they convey. Application of Higher-Order Singular Value Decomposition (HOSVD) on this tensor enables us to distill eigen tensors, representing the principal components of information. These components are then weighted according to their corresponding eigenvalues, facilitating a nuanced quantification of each source's contribution to the collective information structure. Our theoretical analysis, complemented by empirical experiments on datasets from Twitter and various news categories, demonstrates the effectiveness of our approach. The results signify an innovative discovery in modeling and uncovering the deep connections among information sources, paving the way for future research in trustworthy information dissemination and retrieval. The code and dataset utilized in our study are provided in the supplementary materials.

1 Introduction

In the era of digital communication, the task of trustworthy source selection has emerged as a cornerstone for ensuring the integrity of content in social media, retrieval tasks, and even within the domain of Large Language Models (LLMs) employing Retrieval-Augmented Generation techniques (Lewis et al., 2020; Shu et al., 2017). The proliferation of fake news, rumors, and self-media content has underscored the vulnerability of public discourse to widespread misinformation. Such phenomena pose significant risks, capable of misleading public opinion, influencing political elections, fostering hate, and skewing decision-making processes, thereby engendering long-term societal problems (Allcott and Gentzkow, 2017). Against this backdrop, devising an accurate, interpretable, and responsible approach for large-scale trustworthy source selection becomes not only a critical challenge for the NLP and AI communities but also holds immense commercial value (Lecheler and Kruikemeier, 2016).

Despite the recognized need for reliable source verification, existing methodologies often fail to address the multifaceted nature of misinformation (Ekka, 2023). Traditional approaches largely rely on surface-level semantic analysis or fact-checking against known databases. For instance, (Guo et al., 2023) proposes a multiscale transformed model to detect fake news. While useful, these methods cannot fully capture the subtleties of misinformation spread across dynamically evolving social networks (Farrell et al., 2019). Furthermore, they struggle with the scalability required to process the vast amounts of data generated daily on digital platforms and often lack interpretability. These limitations highlight a critical gap in our ability to not only discern but also quantify the reliability of sources in a manner that is both scalable and interpretable. This gap is particularly pronounced in the context of automated systems, such as Large Language Models (LLMs), which utilize retrieval to augment answers, thereby attempting to solve the problem of hallucination where the ability to autonomously assess source trustworthiness can sig-

nificantly enhance the quality and reliability of generated content (Zhang et al., 2023). Consequently, there is a pressing need for innovative approaches that leverage advanced computational techniques to model the intricate relationships between information sources and their content, offering a deeper, more nuanced understanding of source trustworthiness.

Recent advancements in tackling misinformation and source verification have explored a variety of methodologies, from machine learning-based fake news detection to ensuring content integrity. Notably, efforts such as the application of multimodal analysis techniques for semantic inconsistency identification have shown promise (McCrae et al., 2022). Additionally, graph-based models analyzing the network of source interactions have been pivotal in understanding the spread of information, with the potential to track rumor sources (Shelke and Attar, 2019). Moreover, multi-aspect analyses, including entity, sentiment, and text content with BERT-based approaches, have significantly contributed to extracting trustworthy sources (Sufi and Alsulami, 2021). Despite these advancements, the challenge remains to integrate these disparate methods into a cohesive, scalable framework that can operate effectively within the vast, dynamic landscape of social media and digital communication platforms. This integration is crucial for developing an automated, reliable system capable of discerning trustworthy sources in real time. Our research addresses this challenge by leveraging Higher-order Singular Value Decomposition (HOSVD) to model and analyze the complex relationships inherent in digital content dissemination, coupled with a named entity recognition approach to extract key information needs.

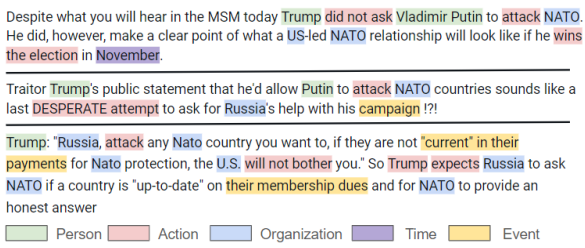

Figure 1: Example of Key Entity Extraction applied to a text excerpt discussing "Trump encouraged Russia to attack NATO".

We have enhanced the BERT-BiLSTM-CRF framework for Named Entity Recognition (NER) by incorporating topic embeddings into the classification layer and specifically annotating entities related to events (Dai et al., 2019). Concurrently, entities deemed irrelevant are labeled as 'O' to streamline the extraction of key entities pivotal for identifying trustworthy sources. An illustration of this refined key entity extraction is presented in Figure 1. This example underscores our focused extraction of entities pertinent to the discussed topic: 'Trump encouraged Russia to attack NATO'.

After extracting key entities, we leverage this information to discern trustworthy sources. This process is governed by three pivotal criteria that are essential for identifying reliable sources, as outlined by (Kozma, 1991):

Spreadability: Trustworthy sources are often widely cited or referred to across various platforms, indicating their acceptance and recognition within the community.

Traceability: A hallmark of reliable sources is the ability to trace key pieces of information back to other independent sources that employ varied methodologies for information collection. This criterion underscores the importance of cross-verification and the presence of a source's information across a diverse set of collection points, enhancing its credibility.

Informative Value: Trustworthy sources are typically rich in detail, providing comprehensive information and evidence to support their claims. The depth and breadth of information offered signify the source's commitment to thoroughness and accuracy, distinguishing it from less reliable counterparts.

In our approach, we consider established criteria to evaluate the trustworthiness of sources, drawing comparisons with prior methodologies for source validation. Pennycook et al. (2018) (Pennycook and Rand, 2018) employed crowdsourcing to identify frequently cited information sources. However, as Figure 2 illustrates, we observe that certain information may be replicated extensively without necessarily contributing to the core narrative, as evidenced by sources S1 and S2. Our analysis also reveals that the absence of certain entities from a source does not preclude credibility; instead, cross-validation with other sources that corroborate those entities can enhance trustworthiness. These insights underpin the development of our model, which facilitates cross-validation and the extraction of independent components that consider their respective contributions. Drawing inspiration from Principal Component Analysis, we utilize a valida-

tion tensor to model this cross-validation process and apply Higher-Order Singular Value Decomposition (HOSVD) to isolate the most significant independent components, subsequently tracing the most contributory sources. This methodology paves the way for a novel and robust method of extracting trustworthy sources.

In summary, this paper makes three principal contributions to the field:

1. We introduce a joint NLP and tensor analysis approach to extract trustworthy sources, combining insights from source topology and semantic content, offering a comprehensive and logical framework for evaluation.

2. We leverage HOSVD to analyze the constructed tensor, providing an interpretable, elegant, and effective alternative to previous methodologies.

3. We validate our model using publicly available Twitter (Zhang et al., 2018) and IFND news (Sharma and Garg, 2023) datasets. Moreover, we employ large language models and human annotation to refine and augment a new dataset, facilitating reproducibility and furthering community research.

Source	Person	Event	Organization	Action	Time
S1	Taylor Swift, Travis Kelce, Hillary Clinton	Super Bowl champions	Kansas City Chiefs	disses	February 11, 2024
S2	Taylor Swift, Travis Kelce, Hillary Clinton	Super Bowl champions	Kansas City Chiefs	disses	February 12, 2024
S3	Taylor Swift, Hillary Clinton	Super Bowl champions	Democrats	Disrespect, called	February 12, 2024
S4	Taylor Swift, Hillary Clinton, Travis Kelce	Super Bowl champions	Kansas City Chiefs	Congratulations to	February 11, 2024

Figure 2: Illustrative example of cross-validation and potential duplication across different sources. The table highlights the overlap and discrepancies in reporting on individuals, events, organizations, actions, and times, suggesting areas for cross-referencing and verification.

2 Related Work

2.1 Name Entity Recognition

Named Entity Recognition (NER) is a critical component in the domain of information extraction, tasked with the identification and categorization of entities like persons, organizations, and locations into predefined groups. Traditional NER models have capitalized on objective functions such as cross-entropy and Conditional Random Fields (CRFs), flourishing under conditions of extensive and balanced annotated datasets (Nguyen et al., 2023).

The NER landscape has undergone substantial evolution. Methods employing weak supervision amalgamate labeling functions with hidden Markov models, facilitating NER model training without fully labeled data (Lison et al., 2020). Additionally, the adoption of graph-based techniques, influenced by dependency parsing, has advanced the detection of complex entity structures, thereby elevating performance benchmarks (Zhao et al., 2021).

In recent advancements, the SpanNER framework treats NER as a span prediction challenge, offering an alternative to the traditional sequence labeling paradigm (Fu et al., 2021). Furthermore, the Unified Machine Reading Comprehension (MRC) framework reframes NER tasks within the context of machine reading comprehension (Li et al., 2019). Techniques for better modeling partially labeled training data are also gaining traction, addressing the challenges of incomplete annotations (Jie et al., 2019). Moreover, distant supervision, particularly through positive-unlabeled learning, has shown potential in generating labeled data for NER (Peng et al., 2019).

For our work, we have chosen the BERT-BiLSTM-CRF architecture as a foundational model for key entity extraction. This decision is informed by the model's exceptional performance and widespread implementation in major libraries such as Huggingface and PySpark (Wolf et al., 2019). Our specialized system is engineered to filter entities relevant to specified topics, thereby enhancing the precision of information extraction and source trustworthiness assessment.

2.2 Higher Order Singular value decomposition

The essence of Higher-Order Singular Value Decomposition (HOSVD) lies in extending the SVD concept from matrices to tensors, enabling a structured analysis of multi-dimensional data (Zare et al., 2018). We commence with the fundamental construct of HOSVD:

Definition (*Tensor and Its n-Mode Matricization*): A tensor $A \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ is represented by its n-mode matricization, $A^{(n)}$, which reorganizes its elements into a matrix format, facilitating linear operations. The crucial transformation is defined as:

$$A^{(n)} \in \mathbb{R}^{I_n \times (I_{n+1} \cdots I_N I_1 \cdots I_{n-1})}, \quad (1)$$

where the organization of elements adheres to a systematic ordering, preserving the tensor's multi-dimensional relationships in a two-dimensional array.

Core Tensor and Singular Vectors: At the heart of HOSVD, the core tensor S and matrices $U^{(n)}$ encapsulate the tensor A's structure:

$$A = S \times_1 U^{(1)} \times_2 U^{(2)} \cdots \times_N U^{(N)}, \quad (2)$$

where $U^{(n)}$ are orthogonal matrices, implying that columns within each $U^{(n)}$ are orthonormal vectors. These matrices reveal the principal directions along each mode, while S captures the magnitude of interactions across these modes.

Singular Values and Orthogonality: The singular values, derived from the Frobenius norms of the core tensor's sub-tensors, quantify the contribution of each mode's singular vectors to the tensor's overall structure. They serve as a measure of data variance along each principal direction:

$$\sigma_i^{(n)} = \|S_{i_n=i}\|_F, \quad (3)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Orthogonal subtensors in S ensure that each dimension contributes uniquely, allowing for a decomposition that highlights the inherent multi-dimensional patterns within A.

Contribution to Structure: HOSVD's ability to disentangle the tensor's complex structure into interpretable components—singular vectors and values—offers profound insights into the data's underlying dimensions. This decomposition not only facilitates dimensionality reduction but also enhances our understanding of the data by isolating its most significant features.

In essence, HOSVD provides a robust framework for analyzing tensors, offering elegant solutions to multi-dimensional data challenges through the mathematical precision of singular values and the interpretability of orthogonal tensors.

3 Methodology

In this section, we detail our approach, beginning with the development of a topic-sensitive Named Entity Recognition (NER) model. This model is trained on a curated dataset that is specifically annotated with entities relevant to the topic of interest, incorporating topic embeddings into the classification layer to enhance specificity. Following this, we describe the construction of our tensor, which is designed to encapsulate the cross-validation relationships among disparate sources. Finally, we illustrate the application of Higher-Order Singular Value Decomposition (HOSVD) and outline the algorithm we have developed to effectively distill trustworthy sources from the tensor representation.

3.1 Topic-Related Named Entity Recognition Model

Our enhanced Named Entity Recognition (NER) model innovatively augments the BERT framework with topic-specific embeddings, thereby enriching the representation of entities in direct relation to a given subject matter. The model is presented in Figure 3 This augmentation is meticulously defined as follows: The BERT architecture generates a

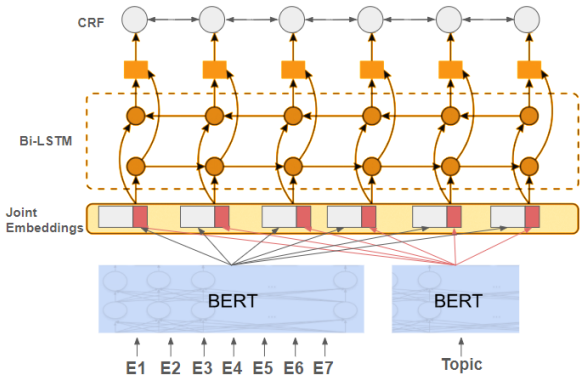

Figure 3: Illustration of the topic-related NER model architecture.

sequence of contextual embeddings:

$$\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}, \quad (4)$$

where $\mathbf{e}_i \in \mathbb{R}^d$ denotes the d-dimensional embedding vector for the i-th token in the sequence, as produced by BERT.

Incorporating topic information, we procure a distinct embedding vector $\mathbf{e}_T \in \mathbb{R}^d$ corresponding to the topic T. This vector is concatenated with each token embedding to form an augmented representation:

$$\mathbf{E}' = \{[\mathbf{e}_1; \mathbf{e}_T], [\mathbf{e}_2; \mathbf{e}_T], \ldots, [\mathbf{e}_n; \mathbf{e}_T]\}, \quad (5)$$

where $[\mathbf{e}_i; \mathbf{e}_T] \in \mathbb{R}^{2d}$ signifies the concatenation of the token and topic embeddings, yielding an enriched feature set for each token.

The sequence of augmented embeddings $\mathbf{E}'$ is then sequentially processed by a Bidirectional

LSTM (Bi-LSTM) network, designed to capture the contextual dependencies:

$$\mathbf{H} = \text{Bi-LSTM}(\mathbf{E}'), \tag{6}$$

where $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$ with each $\mathbf{h}_i \in \mathbb{R}^{2d'}$ represents the hidden states from both forward and backward LSTMs, encapsulating the contextual information from both preceding and succeeding tokens.

These hidden states are subsequently utilized by a Conditional Random Field (CRF) to compute the most likely sequence of tags for named entities:

$$\mathbf{Y} = \text{CRF}(\mathbf{H}), \tag{7}$$

where $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$ is the predicted sequence of entity tags.

The training of our model is formulated as an optimization problem, seeking to maximize the conditional log-likelihood of the true tag sequence given the hidden states:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log p(y_i|\mathbf{H}; \theta), \tag{8}$$

where θ denotes the set of all trainable parameters in the model, and the CRF layer enforces valid transitions between output tags, adhering to the tag sequence constraints.

3.2 Cross-validation Tensor

In the pursuit of accurate information retrieval, our objective is to determine the set of sources that most credibly support the facts regarding N entities within a specific topic. To construct this tensor $\mathcal{T} \in \mathbb{R}^{X \times X \times M}$, we model the cross-validation network among various sources. We include N primary entities of interest and P auxiliary entities, which serve to enhance the cross-validation process—not only within the set of primary entities but also extending to additional validating entities, thus totaling $M = N + P$ entities. The tensor possesses distinctive attributes:

- **Symmetry**: For each entity k, the tensor exhibits symmetry such that $\mathcal{T}_{i,j,k} = \mathcal{T}_{j,i,k}$. This reflects the bidirectional and equivalently weighted relationships between any two sources i and j.
- **Similarity Measure**: The elements of $\mathcal{T}$ are quantified by cosine similarity, thus providing a measure of likeness between the entities reported by different sources.

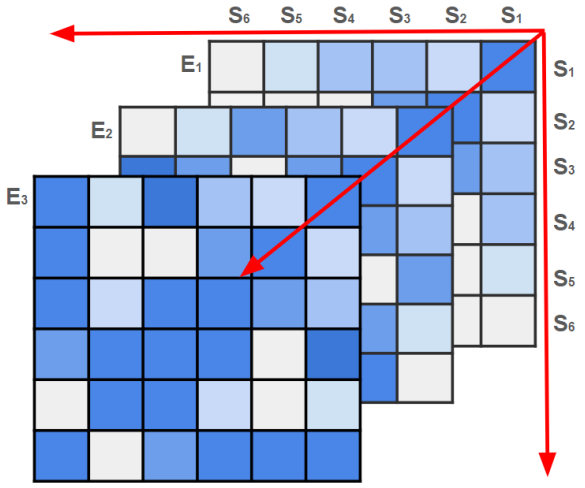

Figure 4: Visualization of the cross-validation tensor. Each layer represents a matrix of sources (S1 to S6) cross-validated with entities (E1 to E3), where the shading intensity reflects the degree of validation between sources.

3.3 Higher-Order Singular Value Decomposition (HOSVD)

The HOSVD of tensor $\mathcal{T}$ decomposes it as:

$$\mathcal{T} \approx \mathcal{S} \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}, \tag{9}$$

where $\mathcal{S}$ is the core tensor signifying the multidimensional interactions across modes. The orthogonal matrices $U^{(1)}$ and $U^{(2)}$ correspond to source dimensions, representing "eigen-sources," and $U^{(3)}$ aligns with the entity dimension, encapsulating "eigen-entities." The orthogonality of the columns within each $U^{(n)}$ ensures the independent capture of variance across dimensions.

3.4 Identifying Source Importance

We quantify source importance based on its contribution to the orthogonal components in $U^{(1)}$ and $U^{(2)}$:

$$Importance(x) = \sum_{i,j,k} \mathcal{S}_{ijk} \cdot (U^{(1)}_{ix} \cdot U^{(2)}_{jx}) \cdot U^{(3)}_{k}, \tag{10}$$

with the summation indicating the total significance of source x. The entity importance is similarly calculated, emphasizing its multidimensional impact:

$$Importance(e) = \sum_{i,j} \mathcal{S}_{i,j,e} \cdot \left(\sum_{k} |U^{(3)}_{ek}| \right), \tag{11}$$

where the aggregated contribution of entity e across all source interactions is considered.

3.5 Ranking and Selection of Sources

Based on the computed importance values, we systematically rank both entities and sources in descending order of their importance. Our algorithm selects sources that maximize coverage of the most important entities, detailed as follows:

1. Initialize the selection set S to hold the selected sources and the set E to track entities that have not yet been covered.

2. Start with the entity of highest importance and identify the highest importance source in N that covers this entity.

3. Incorporate the identified source into the set S and mark all entities covered by this source in the set E as covered.

4. Remove the covered entities from set E and proceed to the next highest importance entity that remains uncovered.

5. Continue iterating through the entities ranked by importance until every entity in N is covered by at least one source in S.

This iterative and greedy approach ensures that at each step, we select a source that covers the most important entity not yet accounted for, ultimately resulting in a set of sources S that collectively covers all entities in N comprehensively and efficiently. The algorithm terminates when there are no uncovered entities left in N, thereby achieving complete coverage with an optimally curated subset of sources.

4 Experiments

In this section, we empirically assess our proposed methodology against three pivotal research questions:

1. *Effectiveness in Trustworthy Source Extraction*: How does our approach compare to previous methods in accurately extracting trustworthy sources?

2. *Scalability*: Is our approach capable of handling large-scale scenarios where the number of sources increases significantly?

3. *Impact of Entity Importance*: Through an ablation study, we examine the effect of varying the number of entities used for cross-validation. Entities are incrementally excluded based on their assessed importance to evaluate the impact on the accuracy of source trustworthiness assessment.

Our experiments are designed to rigorously test these aspects, thereby establishing the efficacy and robustness of our tensor-based model for identifying reliable sources in an information-rich environment.

4.1 Dataset

Our experiments utilize datasets predominantly sourced from Twitter and various news outlets, all of which are publicly accessible. To enrich these datasets and ensure their relevance to current events, we have also conducted additional data collection:

1. **Twitter (X) Dataset** (Ma et al., 2017): Originally created for rumor detection, the Twitter dataset comprises two subsets: Twitter15 and Twitter16. These subsets were initially designed for binary classification tasks—distinguishing between rumors and non-rumors related to specific events. The datasets initially included 276,663 and 173,487 users, respectively. Through a process of data augmentation, which involved additional Twitter data crawling and entity tagging, we have expanded the dataset to encompass 2,234,847 tweets. The augmented dataset covers 3,326 topics and extracts 6 primary entities (Person, Organization, Time, Action, etc.) along with 10 auxiliary entities (Color, Age, Size, etc.).

2. **News Dataset** (Sharma and Garg, 2023): This dataset, primarily used for fake news detection, contains a collection of news items ranging from the year 2013 to 2021. It includes 37,809 verified news articles and 19,059 articles identified as fake news. For the purposes of our study, we have retained only the textual content. Further expansion was achieved by extracting additional news items using Google based on the topics present in the dataset. Consequently, our augmented dataset now includes 1,872,674 sources across 20,332 topics, with the same set of 6 key entities and 10 additional auxiliary entities identified.

This comprehensive dataset provides a robust foundation for our study, offering a wide range of topics and entities to analyze source trustworthiness in the context of information dissemination on

social media and news platforms. The preprocessed and augmented dataset will be made available upon completion of the review process at this link.

4.2 Baselines

To benchmark the performance of our approach, we juxtapose it with a suite of baseline methodologies, each representing the forefront of trustworthy source identification:

1. **Bi-GCN** (Bian et al., 2020): Utilizes a bi-directional graph convolutional network to model rumor propagation dynamics.

2. **PLAN** (Khoo et al., 2020): Employs a transformer-based attention model for analyzing tweet interactions.

3. **TD-RvNN** (Ma et al., 2018): Adopts recursive neural networks to mirror the hierarchical spread of tweets.

4. **FNDNet** (Kaliyar et al., 2020): Implements a deep CNN for automatic feature extraction in fake news detection.

5. **FND-NS** (Raza and Ding, 2022): Leverages a Transformer architecture for predictive modeling in fake news analysis.

6. **FakeBERT** (Kaliyar et al., 2021): Combines BERT with parallel blocks of a single-layer CNN, each with varying kernel sizes and filters, to address the complexities of natural language understanding and the inherent ambiguity in fake news detection.

7. **TSE-HOSVD**: Our proposed approach employs a novel tensor-based method to extract and verify trustworthy sources using higher-order singular value decomposition.

These baselines span a range of techniques, from graph-based models to deep learning architectures, and provide a comprehensive backdrop against which the superiority of our proposed method can be measured.

4.3 Evaluation Metrics

Distinct from prior studies focused on fake news or rumor detection—typically framed as classification problems—our research emphasizes the retrieval of the most trustworthy sources. We evaluate sources that achieve the highest trust scores and encompass all critical entities, ensuring their matches are accurate. This process involves verifying the correctness of key entities within these sources. To measure the efficacy of our approach, we employ precision, recall, and F1 score metrics.

4.4 Experimental Settings

We utilize a 5-fold cross-validation method to optimize and train our baseline models, aligning our experimental design with scenarios likely to be encountered in real-world deployments. To support this, we incorporate a vector retrieval strategy, adopting Faiss (Johnson et al., 2019) for the efficient handling of baseline model embeddings, pre-computed prior to topic identification. This approach enables us to conduct testing on a large scale effectively. The computational infrastructure supporting our model's training and execution includes four NVIDIA Tesla P100 GPUs and 128 vCPUs provided by Google Cloud Services. All models are implemented using PyTorch. We use GPU -based truncate HOSVD to process our approach, which is an advanced approach to approximate HOSVD efficiently.

4.5 Results

We present our experimental results on the Twitter and News datasets in Tables 1 and 2, respectively. The highest scores are in bold, and asterisks (*) denote statistical significance over the second-best results. Our approach achieved a 5.5% improvement over the state-of-the-art PLAN model on the Twitter dataset, and a 3.4% increase on the News dataset, both statistically significant. The results indicate that models considering network topology, such as ours and Bi-GCN, excel on social media datasets. In contrast, content analysis models like FakeBERT perform better on news content. However, our TSE-HOSVD model, which integrates both topology and content analysis, sets a new state-of-the-art across both datasets. This underscores the importance of combining these factors for reliable source extraction.

The scalability test results for processing time as the number of sources increases are depicted in Figure 5. Initially, for a smaller number of sources, baseline models outperform TSE-HOSVD in terms of speed. However, as the source count escalates—a typical scenario in real-world applications—the inference time for these models increases significantly. Since only inference time is considered here, reflecting a more realistic usage scenario, TSE-HOSVD proves to be more effi-

Approach	Recall	Precision	F1 Score
Bi-GCN	0.819	0.827	0.823
PLAN	0.854	0.893	0.873
TD-RvNN	0.753	0.781	0.767
FNDNet	0.817	0.822	0.819
FND-NS	0.837	0.865	0.851
FakeBERT	0.846	0.871	0.858
TSE-HOSVD	**0.897***	**0.946***	**0.921***

Table 1: Comparative Performance of Baselines on the Twitter (X) Dataset.

Approach	Recall	Precision	F1 Score
Bi-GCN	0.786	0.808	0.797
PLAN	0.861	0.883	0.872
TD-RvNN	0.791	0.811	0.801
FNDNet	0.843	0.859	0.851
FND-NS	0.850	0.876	0.863
FakeBERT	0.874	0.904	0.889
TSE-HOSVD	**0.903***	**0.935***	**0.919***

Table 2: Comparative Performance of Baselines on the News Dataset.

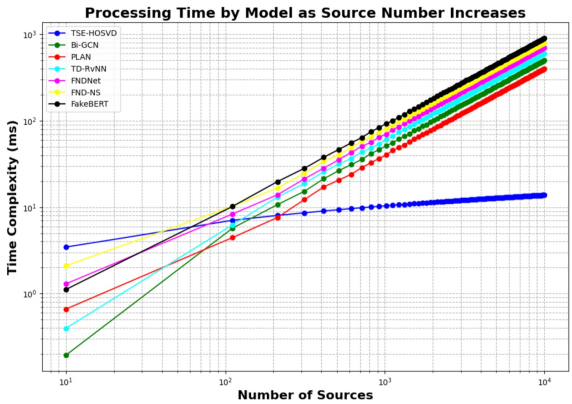

Figure 5: Processing Time by different baseline models as Source Number Increases on Twitter (X) dataset.

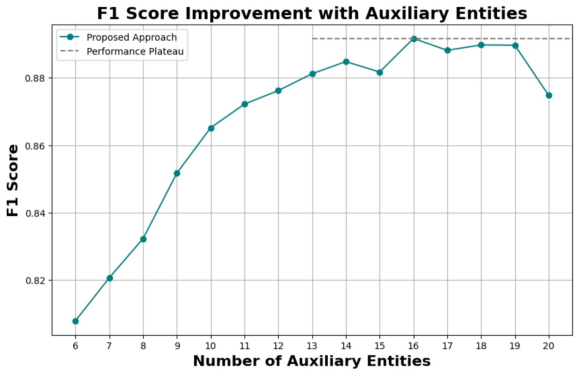

Figure 6: The impact of auxiliary entities on trustworthy detection on Twitter (X) dataset.

cient, especially when handling over 1,000 sources. This underscores the proposed model's advantage in large-scale environments, as demonstrated by its comparatively lower time complexity and enhanced scalability.

4.6 Ablation Study

Our ablation study, depicted in Figure 6, assesses the impact of auxiliary entities on model performance. The results indicate that auxiliary entities can indeed improve the F1 scores, signifying enhanced detection of trustworthy sources. Nonetheless, the study reveals that there is an optimal number of entities for cross-validation; incorporating too many can dilute the focus on primary attributes critical to understanding the topic, potentially lowering performance. This threshold effect is evident as performance plateaus when the total number of entities reaches 16, informing our choice of this as the optimal entity count. Entity importance, calculated as per Equation 11, guides the ranking and selection process in this optimization.

5 Conclusion

In this work, we have introduced a novel approach for trustworthy source extraction, a critical aspect of information retrieval, natural language processing, and retrieval-augmented generation tasks. By integrating a topic-sensitive NER technique with HOSVD, we have established a robust method for discerning reliable information sources. The topic-sensitive NER effectively parses content, and HOSVD analyzes the underlying network topology. Together, they yield state-of-the-art results, surpassing previous methodologies. The application of HOSVD provides a nuanced understanding of source validation akin to the role PCA plays in feature extraction for images. Our experimental results demonstrate that our approach not only achieves state-of-the-art accuracy but also excels in scalability, making it well-suited for large-scale applications.

6 Ethics Consideration

In the development and deployment of models for trustworthy source extraction, ethical considerations must be at the forefront to ensure that the technology is used responsibly and does not perpetuate bias or misinformation.

Firstly, the potential for the model to inadver-

tently filter out minority voices or critical perspectives that deviate from the majority must be acknowledged. This can lead to a reinforcement of dominant narratives and the marginalization of underrepresented groups. It is imperative to continuously evaluate and adjust the model to account for a diverse range of sources and viewpoints, maintaining a balance between trustworthiness and inclusivity.

Secondly, the use of topic-sensitive Named Entity Recognition (NER) may raise privacy concerns. Entities often include personal names and locations, and their extraction and analysis must be handled with respect to individuals' privacy rights. Proper anonymization techniques and adherence to data protection regulations, such as the General Data Protection Regulation (GDPR), are essential.

Additionally, the system's role in content moderation has implications for freedom of expression. The criteria used to assess trustworthiness must be transparent to avoid censorship and ensure that the suppression of content is justified and non-discriminatory.

Furthermore, the reliance on automated systems to determine the credibility of sources can lead to overtrust in technology and underappreciation of human critical thinking skills. It is crucial to emphasize that the model serves as an aid rather than a replacement for human judgment.

Finally, the deployment of such models across different cultural and linguistic contexts requires careful consideration. What is considered trustworthy in one culture may not be seen the same way in another, and models must be adapted accordingly to avoid cultural insensitivity or bias.

As we continue to refine our model, we must engage with these ethical concerns actively and involve a range of stakeholders, including ethicists, community representatives, and domain experts, to guide responsible research and application.

7 Limitations

The methodology introduced in this study, while effective, is subject to certain limitations that offer avenues for refinement. The current approach relies on the precision of topic-sensitive Named Entity Recognition (NER), which can be susceptible to the nuances and variability of language across different contexts and datasets. In cases where NER performance is suboptimal, the overall efficacy of source extraction may be affected.

Additionally, the model's performance is contingent on the quality and representativeness of the auxiliary entities. An overemphasis on auxiliary entities that do not contribute meaningful information can lead to a dilution of performance, as observed in the ablation study.

While the model showcases scalability, the increase in processing time with the number of sources remains a concern for real-world applications where source numbers can be extraordinarily high. Therefore, optimizing the model to handle vast source numbers with minimal computational overhead is a critical consideration for future enhancements.

In addressing these limitations, subsequent research could focus on improving the robustness of topic-sensitive NER, enhancing the selection criteria for auxiliary entities, and optimizing the computational aspects of HOSVD. The goal will be to refine the model to ensure consistent performance across a wider array of scenarios, bolstering its practicality for deployment in diverse and dynamic information landscapes.

References

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

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556.

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using bert bilstm crf for chinese electronic health records. In *2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei)*, pages 1–5. IEEE.

Bosco Ekka. 2023. A comprehensive analysis of evaluation strategies for online information truthfulness.

Justin Farrell, Kathryn McConnell, and Robert Brulle. 2019. Evidence-based strategies to combat scientific misinformation. *Nature climate change*, 9(3):191–195.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. Spanner: Named entity re-/recognition as span prediction. *arXiv preprint arXiv:2106.00641*.

Zhiwei Guo, Qin Zhang, Feng Ding, Xiaogang Zhu, and Keping Yu. 2023. A novel fake news detection model for context of mixed languages through multiscale

transformer. *IEEE Transactions on Computational Social Systems*.

Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. Better modeling of incomplete annotations for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 729–734.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

Rohit Kumar Kaliyar, Anurag Goswami, Pratik Narang, and Soumendu Sinha. 2020. Fndnet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, 61:32–44.

Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. 2020. Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8783–8790.

Robert B Kozma. 1991. Learning with media. *Review of educational research*, 61(2):179–211.

Sophie Lecheler and Sanne Kruikemeier. 2016. Re-evaluating journalistic routines in a digital age: A review of research on the use of online sources. *New media & society*, 18(1):156–171.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019. A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*.

Pierre Lison, Aliaksandr Hubin, Jeremy Barnes, and Samia Touileb. 2020. Named entity recognition without labelled data: A weak supervision approach. *arXiv preprint arXiv:2004.14723*.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

Scott McCrae, Kehan Wang, and Avideh Zakhor. 2022. Multi-modal semantic inconsistency detection in social media news posts. In *International Conference on Multimedia Modeling*, pages 331–343. Springer.

Ngoc Dang Nguyen, Wei Tan, Lan Du, Wray Buntine, Richard Beare, and Changyou Chen. 2023. Auc maximization for low-resource named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13389–13399.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. *arXiv preprint arXiv:1906.01378*.

Gordon Pennycook and David G Rand. 2018. Crowdsourcing judgments of news source quality. *SSRN. com*.

Shaina Raza and Chen Ding. 2022. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics*, 13(4):335–362.

Dilip Kumar Sharma and Sonal Garg. 2023. Ifnd: a benchmark dataset for fake news detection. *Complex & Intelligent Systems*, 9(3):2843–2863.

Sushila Shelke and Vahida Attar. 2019. Source detection of rumor in social network–a review. *Online Social Networks and Media*, 9:30–42.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Fahim K Sufi and Musleh Alsulami. 2021. Automated multidimensional analysis of global events with entity detection, sentiment analysis and anomaly detection. *IEEE Access*, 9:152449–152460.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Ali Zare, Alp Ozdemir, Mark A Iwen, and Selin Aviyente. 2018. Extension of pca to higher order data structures: An introduction to tensors, tensor decompositions, and tensor pca. *Proceedings of the IEEE*, 106(8):1341–1358.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC*

2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15, pages 745–760. Springer.

Xinyan Zhao, Haibo Ding, and Zhe Feng. 2021. Glara: Graph-based labeling rule augmentation for weakly supervised named entity recognition. *arXiv preprint arXiv:2104.06230*.