Wiki Workshop (11th edition) – June 20, 2024

Comparative Analysis of Knowledge Graphs Constructed from Fake News and Legitimate News Sources

Shivam Kumar and **Dwaipayan Roy**
Indian Institute of Science Education and Research Kolkata

Abstract

This paper explores the utility of Wikipedia for distinguishing real and fake news articles through Knowledge Graph (KG) analysis. We compare the KGs constructed from real and fake news datasets, leveraging Wikipedia and Named Entity Recognition (NER) to extract and normalize entities. Our investigation examines structural differences within these graphs to assess the effectiveness of Wikipedia-based KGs in identifying fake news sources.

Keywords: Knowledge Graph; Comparative Analysis; Wikipedia; Named Entity Recognition; Fake News.

Introduction

The ever-growing prevalence of fake news poses a significant challenge to discerning truth in today's information landscape (Zhou and Zafarani, 2020). This paper presents the potential of knowledge graphs (Ji et al., 2022), a powerful data representation tool that works based on the graphical structure, to differentiate between real and fabricated news articles. While KGs power applications like search engines and recommendation systems, their effectiveness hinges on reliable data sources. If compromised sources create indistinguishable KGs from real and fake news, their credibility for tasks like truth detection becomes a concern.

Our approach hinges on two key techniques: Wikipedia and Named Entity Recognition (NER). NER (Huguet Cabot and Navigli, 2021) acts as the first step, adeptly identifying crucial entities like people, locations, and organizations within the text of news articles. Following this, Wikipedia normalization, this process verifies that entities referencing the same Wikipedia page are represented within the knowledge graph. This structured data forms the foundation of our KGs, where entities become nodes and their relationships are visualized as edges. This underlying structure of "entity-relation-entity" allows for in-depth analysis.

We examine the structural properties of KGs constructed from two distinct datasets – one containing legitimate news and the other harbouring fabricated or misleading (fake) information. This work tackles a critical question: *can knowledge graphs built with NER and Wikipedia normalization effectively distinguish real from fake news sources?* We analyze structural differences between KGs constructed from real and fake news datasets. This investigation aims to illuminate the effectiveness of these techniques in fake news detection. Our research not only offers insights into the nature of real and fake news but also sheds some initial light on the potential and limitations of NER and Wikipedia normalization in this evolving domain.

Hypothesis

We explore whether the knowledge graphs generated from Fake and Real data may exhibit distinct structural characteristics. Our focus is on discerning the credibility of news content through these graphs. We anticipate that genuine news KGs will display extended connections and tighter clustering of entities. In contrast, fake news graphs are likely to appear fragmented, featuring many isolated clusters and shorter links. This hypothesis stems from the assumption that real news typically involves well-established entities with strong inter-relationships, leading to denser network structures within the KG. Conversely, the inherent lack of factual accuracy in fake news translates to fragmented graphs with smaller clusters and weaker connections between entities, as they may be isolated or have fewer established relationships.

Our Method

For our empirical study, we employ the LIAR dataset (Wang, 2017), a widely-used dataset containing statements labelled with their truthfulness. The NER and normalization process involved splitting the text into spans of approximately 128 tokens, equivalent to around 96 English words, with relations extracted from each span. We employed REBEL-LARGE model for doing this. To ensure consistency, entities identified by the NER step such as "Napoleon Bonaparte" and "Napoleon" were normalized using the Wikipedia, confirming if they shared the same Wikipedia page and then normalizing them to the page's title, if applicable. We created *Word Cloud* to visualize the entity and its frequency, filtering out generic non-informative terms beforehand. We chose entities with frequency $>$ 0.25 from the Word Cloud.

This process was applied to both datasets. Relationships involving these common named-entities were selected to construct the KGs. This approach ensures a consistent framework for comparing the structural characteristics of the real and fake news graphs on same topic(s).

Optimizing Wiki Search on Wiki Dump

As illustrated by the figure 1, the primary index, given with WikiDump, comprising of words and corresponding start-byte data is first sorted, then grouped, each containing a defined number of words. Subsequently, a secondary index is constructed using the initial element which acts as a pointer to these grouped segments. This systematic approach ensures the swift retrieval of information. After finding the entity, we use the start-byte data to extract only the relevant portion from the WikiDump.

The processing time with Wikipedia API typically requires approximately 20 seconds to return search results (Wikipedia return results in around 5 seconds, but multiple API calls for normalization process increases the return time). Our optimized method dramatically reduces the processing time to around 1 second (on AMD Ryzen 7 6800H and 40Mbps internet connection).

By efficiently organizing and accessing data, our approach significantly enhances the speed and responsiveness of Wikipedia search queries, making it a valuable asset for applications requiring rapid access to Wikipedia's vast information resources.

Results

Upon visualizing the knowledge graphs generated from both datasets, it became evident that both datasets exhibited the generation of similar graph structures. Each dataset contained one to two dense clusters with long links, along with smaller clusters characterized by shorter links. This unexpected similarity challenges our initial assumptions regarding the expected differences between the graphs of Real and Fake. Building KGs from fake and real news sources yielded surprisingly similar structures. Figure 2 depicts the distribution of node degrees in both knowledge graphs. It reveals a strikingly similar pattern, with a high concentration of low-degree nodes and a sharp decline as degree values increase. Examining portions of the KGs (Figures 3 and 4) unearthed an unexpected similarity between real and fake news data. Both displayed a comparable structure of one or two central clusters with longer links, surrounded by smaller clusters with shorter connections. This unforeseen result challenges our initial belief that real and fake news graphs would manifest distinct characteristics.

Discussion and Conclusion

While it can be anticipated that distinct graph patterns will be formed based on the differing characteristics of real and fake news, the presence of analogous structures suggests other factors may play a more significant role. Possible explanations could involve commonalities in language usage, topical overlap, or shared entities across news articles irrespective of their veracity. Additionally, it is plausible that the NER and normalization process, although designed to capture meaningful entities, may inadvertently introduce biases that homogenize the resulting graphs. Further investigation into these factors is crucial for a deeper understanding of the underlying mechanisms shaping graph formations. Exploring alternative graph construction methodologies or incorporating additional features may provide insights into distinguishing between real and fake news datasets more effectively. Overall, this unexpected discovery underscores the complexity of graph analysis and the need for nuanced approaches in uncovering underlying patterns in diverse datasets.

Our project focused on improving Wikipedia search speed and analyzing real and fake news through KGs. While the search speed improvement is promising, the analysis of KGs revealed unexpected similarities between real and fake news data. This challenges our initial assumptions and necessitates further investigation into potential reasons like shared language, topical overlap, or analysis bias. Future research could explore alternative methods or features to better differentiate real from fake news. Overall, the project contributes to both information access and data analysis understanding. The code implementations are available here for research purposes.

References

[Huguet Cabot and Navigli2021] Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

[Ji et al.2022] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.

[Wang2017] William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proc. of 55th ACL (Volume 2: Short Papers)*, pages 422–426, July.

[Zhou and Zafarani2020] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5), sep.

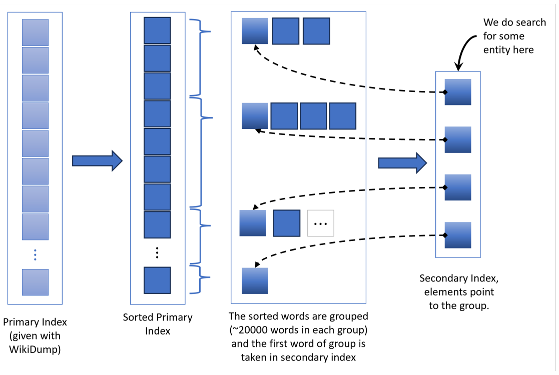
Primary Index (given with WikiDump)
Sorted Primary Index
The sorted words are grouped (~20000 words in each group) and the first word of group is taken in secondary index
We do search for some entity here
Secondary Index, elements point to the group.
...

Figure 1: A simplified representation of data structure made for optimising Search for effectively searching and retrieving data from a local WikiDump which involves organizing data into a primary index, sorting it, and grouping it for efficient management. A secondary index is created to expedite search operations, utilizing the first word from each group for rapid data access and reference.

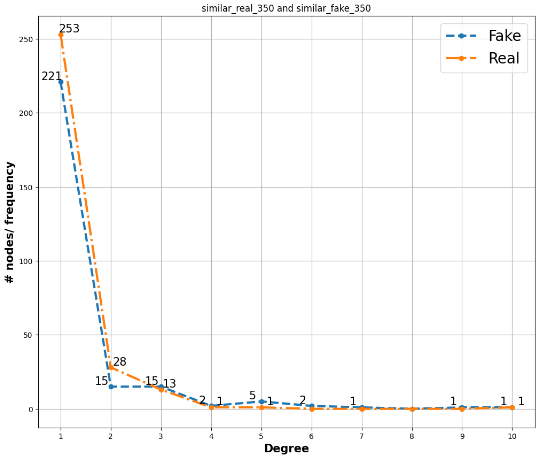

Figure 2: Plot showing the frequency of degree of a node in knowledge graphs created with fake and real information. It demonstrates a comparable distribution pattern between 'Fake' and 'Real' nodes, featuring a prevalent occurrence of nodes with low degrees followed by a sharp decrease as degree values rise. This similarity implies structural resemblance between the two networks in terms of node degree distribution, suggesting the potential challenge of differentiating between 'Fake' and 'Real' solely based on network topology.

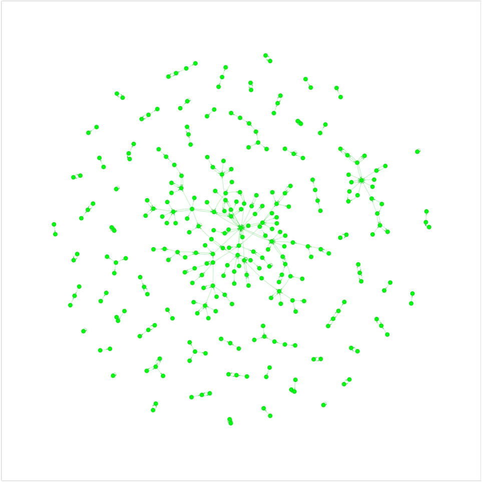

Figure 3: Knowledge Graph of 200 real statements: the graph shows a network with a central cluster, a large, densely connected group, surrounded by smaller, separate peripheral clusters.

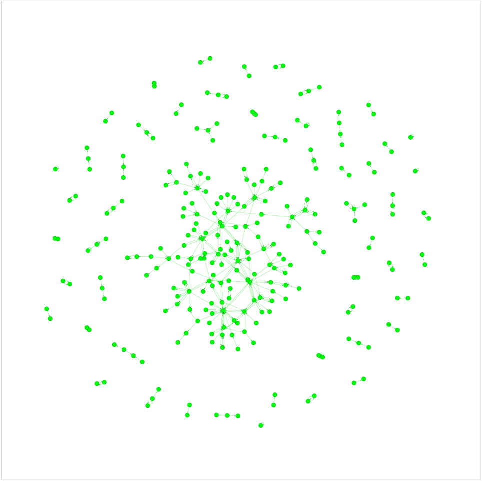

Figure 4: Knowledge Graph of 200 fake statements: the graph shows a network with a central cluster, a large, densely connected group, surrounded by smaller, separate peripheral clusters.