Stories from Blogs: Computational Extraction and Visualization of Narratives

Muhammad Nihal Hussain, Kiran Kumar Bandeli[1], Hayder Al Rubaye, and Nitin Agarwal[2]
[2]Jerry L. Maulden-Entergy Chair Professor of Information Science
University of Arkansas at Little Rock, Little Rock, Arkansas, United States
{mnhussain, hkalrubaye, nxagarwal}@ualr.edu
[1]KiranKumar.Bandeli@walmart.com

Abstract

Social media platforms are designed to serve as an avenue for people to connect with like-minded individuals to discuss their views and promote democracy. However, they have become a tool for deviant actors to undermine the same. Due to the anonymity and perceived less personal risk, deviant groups coordinate on these platforms to spread fake news, misinformation, and disinformation. Social media platforms such as blogs that are unregulated and provide richer space for content generation are strategically used for agenda setting, content framing, and weaponizing narratives to radicalize mobs and provoke hysteria. The recent events and protests coordinated through social media demonstrate the critical need for tools to identify these fringe narratives early on. In this paper, we demonstrate a narrative visualization tool that provides an analyst the ability to identify prominent themes and associated narratives. The tool builds upon published framework to extract narratives from blogs and is available for public use through the Blogtrackers application.

1 Introduction

Social media is characterized as a powerful online interaction and information exchange medium. However, it has given rise to new forms of deviant behaviors such as spreading fake news, misinformation, and disinformation. Due to afforded anonymity and perceived less personal risk of connecting and acting online, deviant groups are becoming increasingly common. These groups harness the power of social media, weaponize narratives to polarize, radicalize and mobilize citizens. More recently, there is a surge in misinformation, conspiracy theories, and scams pertaining to COVID-19. The problem of misinformation is worse than the pandemic itself. Hence, the phenomenon is termed infodemic, or more specifically, misinfodemic [Min18]. Like the pandemic, misinformation narratives are also rising exponentially. These narratives are more difficult to track than the epidemic, as they can originate in the dark corners of the internet. To make matters worse, we cannot enforce lockdown on the Internet to stop the spread of this infodemic. To eliminate these radical effects on social media, it is important to track these misinformation narratives, as they develop, to build counter measures to quickly stem their damage to the society.

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Finlayson (eds.): Proceedings of the Text2Story'21 Workshop, Online, 1-April-2021, published at http://ceur-ws.org

Analysis of narratives on social media and building counter measures requires not only efficient algorithms to accurately extract narratives and also intuitive tools to sift through them. While there have been several studies [MA20], [DVK19] focused on extracting narratives from a corpus, there is a lack of intuitive tools that visualize them. Moreover, with the volume and the velocity with which content is generated on social media, being able to only extract narratives would not be enough. In this paper, we demonstrate a narrative visualization tool for social media, especially blogs. The tool extracts narratives leveraging our published research [BHA20] on narrative extraction and is integrated with the Blogtrackers [Blo18] application (https://btracker.host.ualr.edu) and provides a user a tree-like structure to visualize narratives anchored around prominent keywords or keywords of interest. The tool allows users to provide feedback thereby helping to improve the narrative extraction approach. The proposed tool is scalable, language independent, and adaptive to other narrative extraction approaches. It further offers several customizations, discussed in section 4, to improve the overall user experience.

Rest of the paper is organized as follows. Section 2 discusses several narrative extraction and visualization techniques. In section 3, we briefly explain the narrative extraction framework used in this study. Proposed visualization tool for narratives is explained in section 4. In Section 5, we discuss the challenges and limitations of the tool presented in the paper. We conclude with proposed future direction in section 6.

2 Literature Review

Narrative is defined as "a spoken or written account of connected events" and researchers have identified several approaches to extract the same. Chou et. al. [CHFA11] focused on identifying linguistic and thematic characteristics of a given text and listed common attributes of narratives. Studies by Corman et.al [CRF12] and Ruston [RCS+16] focus on semantic triplets of subject-verb-object to extract narratives. They also discuss the use of cultural aspects in framing of narrative by extremists to influence crowds. Whereas, Holmstrom [Mir] focused on temporal words in the text to explain intent behind multiple interconnected narratives discussed on social media. Dirkson et. al [DVK19] found that combination of temporal ques and the linguistics, especially triplets, characterize narrative better in health related datasets.

While the majority of the studies in this domain focused on extracting narratives, a few focused on visualizing them efficiently. A study by McKenna et. al. [MHRL+17], investigated the elements that affect users' perception in data driven narrative visualization by analyzing 80 well known visual narratives hosted over the web. Authors identified 7 "flow-factors" - navigation input, level of control, navigation progress, story layout, role of visualization, story progression, and navigation feedback, and provided a score mechanism to evaluate the visualization. Their results showed that the flow-factors like visuals and navigation feedback affect engagement more than level of control. In another study [SML18], authors developed COMFRE (COMparing FREquencies) to overcome the drawback in word clouds and provide better comparison of word frequency distributions visually across two corpora by combining slope charts and histogram contours. But it does not display actual word frequencies, has a lot of whitespace and does not allow word frequency comparison with the same corpora. In another study, authors [KNM20] introduced time-sets visualization to allow users to compare multi word occurrences in media content over time. Their results showed improved user experience for smaller word sets and fewer media sources but the proposed visualization became very complex when word sets grew and multiple media sources were introduced making the visualization difficult to comprehend and compare.

3 Research Methodology

To extract the narratives from blogs, we use our previously published method [BHA20] (fig. 1) briefly explained here. For a given set of blogs, we begin by extracting names of prominent personalities or locations and focus on discussions around them. To accomplish this, we use named entities extraction to extract entities and use their document frequency to rank and assess their prominence in the blogs. Later, we input a combination of blog posts and identified entities to the network topic modeling module to identify the topics of discussion associated with the entities. Network topic modeling helps identify the topics of discussion that are exclusive to the entities but also provides us overlapping topics. Example, an analysis of recent events in the USA would provide COVID-19 and Donald Trump as prominent entities. The discussion topics exclusive to COVID-19 would include the overall impact of COVID-19, lockdowns, cure and misinformation about it. Whereas topics exclusive to Donald Trump would be about the elections and other political events. The overlapping topics would discuss Donald Trump's policies toward COVID-19 and any misinformation that includes both entities. The advantage of extracting overlapping themes provides ability to drilldown or customize analysis. The number of topics extracted per entity is decided based on parameters in the LDA (Latent Dirichlet Allocation) model and

tuned to get the optimum number of topics. Parameters in the LDA model are log-likelihood and beta. In LDA, both topic distributions - over documents and words, have correspondent priors, which are denoted typically with alpha and beta, and because these are the parameters of the prior distributions, are commonly referred to as hyper-parameters. For a dataset, the log-likelihood is calculated for each iteration with topics as 1, 5, 10, 25, 50 etc. Ideally, the LDA model gets "better" at describing the data and increases over time. Eventually this value will level off, as subsequent iterations make negligible improvements to the model. On the other hand, low beta value places more weight on having each topic composed of only a few dominant words. Therefore, the combination of these hyper-parameters helps in choosing number of topics.

Once the topics for the entities are extracted, we filter down to blog posts that are most contributing to the topics based on their topic distributions. Later, from these blog posts, we only extract the sentences that mention the entity associated with the topic. We leverage NLP techniques like POS tagging, chunking to extract noun phrases and verb phrases. After this step, we define grammar rules that capture specific patterns that have been empirically identified to extract narratives. Finally, we merge narratives for each entity based on their similarity and rank different narratives based on their dominance in the dataset. Language does not impose barriers given the statistical nature of our narrative extraction approach. However, In this study, we experimented with English blogs only. We intend to apply our methodology to non-English text and evaluate its efficacy.

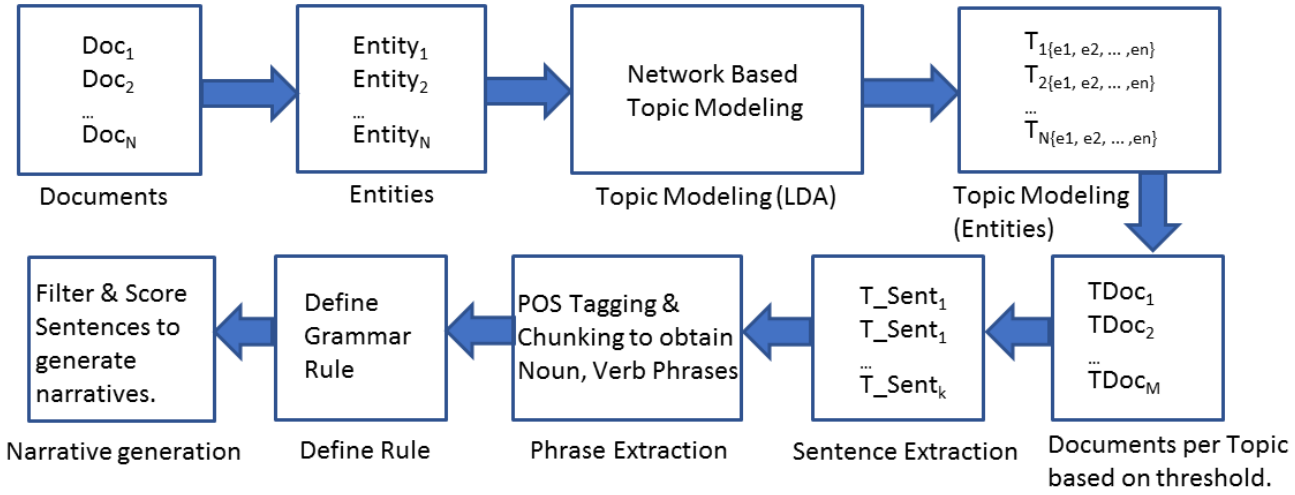

Figure 1: Framework to extract narratives

4 Narrative Visualization

Once the narratives are extracted, they are visualized on the Blogtrackers tool [Blo18] as shown below (fig. 2). The tool lists prominent entities as keywords in a collapsible tree of items where each item contains narratives that are associated with the entity. The ability to collapse keywords and narratives enhances the user experience by allowing the user to focus only the keywords of their interest. Users can interact with each of those keywords to show their related narratives. They can also click on any of the narratives to expand and show a list of its related posts in a horizontal scrollable view (fig. 4). The goal here is to prevent posts from taking a lot of vertical space that might expand the page and urge users to scroll too much.

To enhance user analytical capabilities the tree allows users to customize and search keywords of their own interest. Users can use the search bar to search their own keywords and the tool will automatically list the narratives associated with them. The tool also allows users to combine keywords to view overlapping narratives. This can be accomplished by switching to editing mode where every keyword and narrative becomes editable. Multiple keywords can then be selected and merged together to create collections of subjects. Keywords can also be removed from a collection, and collections can themselves be removed. Overlapping topics among entities provided by network-based topic models enables keyword compounding and extraction of compound narratives (fig. 3).

The tool allows users to provide feedback by editing or improving narratives (fig. 5). Updated narratives are stored which are used later to improve the narratives extraction algorithm. Moreover, the narrative visualization

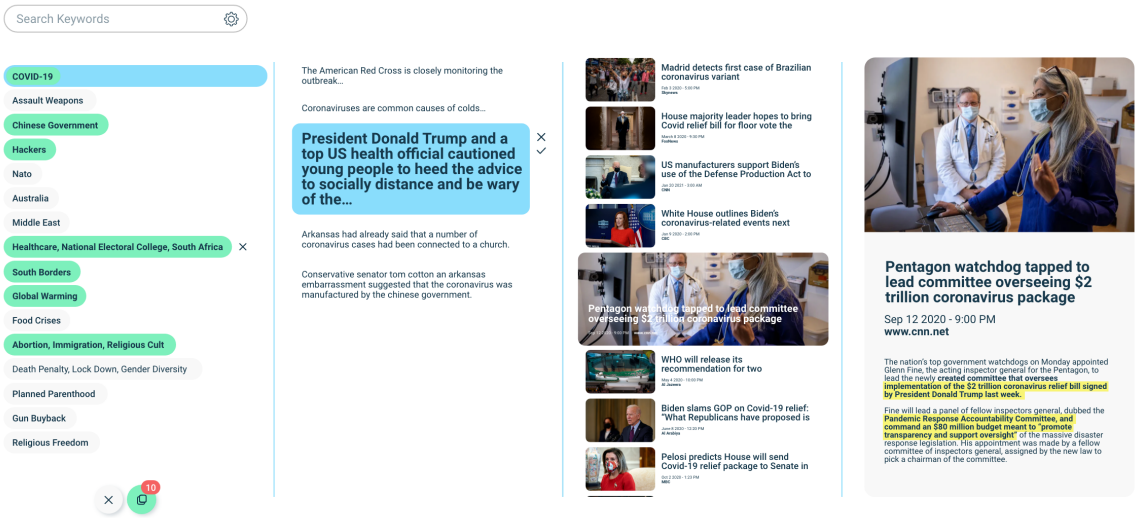

Figure 2: Narrative visualization tool in Blogtrackers application

tool is adaptive to any narrative extraction approach and works in a plug-n-play manner. Additionally, the narratives are pre-computed making the approach scalable to the deluge of data observed from social media platforms.

Finally, to improve user experience the tree was color coded with vibrant, high contrast, and bright colors, where every color has a meaning. Yellow has been chosen to represent non expanded keywords. Blue was picked to represent keywords that are in an expanded state, this would help users differentiate between the two states. Whereas keywords change to Green when the tree is in edit mode, this indicates that the keywords can be selected and edited by the user.

5 Challenges and Limitations

While the narrative extraction algorithm used in the above demonstrated tool identifies narratives with reasonable accuracy, it still has several challenges and limitations. The grammar rule currently used is tailored based on empirical observations and has subject bias. Moreover, the rule could fail for complex sentences and its results might not be generalizable. There is a need to improve the narrative extraction algorithm to make it more generalizable.

The visualization of the narrative, although, provides analysts several customizations to help discover narratives associated with their keyword(s) of interest, it has few limitations. The UI has only been tested by analysts and developers that had prior knowledge of the Blogtrackers tool [Blo18]. The UI needs to be evaluated by a wider audience to assess its usability.

6 Conclusion and Future Work

With social media affecting almost every aspect of our life and with the recent events that demonstrated the power of social media to radicalize and mobilize crowds, there is a critical need for tools to sift through social media to assist authorities to identify extreme or fringe narratives. In this paper, we demonstrated a tool to effectively visualize narratives. The tool builds upon existing narrative extraction algorithm [BHA20] and inherits its strengths such as accuracy, language independence, scalability, but also its limitations. The grammar rule, a core aspect, of the algorithm is non-generalizable and needs to be improved to handle complex sentences. The tool designed is independent of narrative extraction algorithms and can automatically adapt to another algorithm. The tool provides capabilities to identify simple and compound narratives. It also allows users to provide feedback which is used in improving the narrative extraction algorithm.

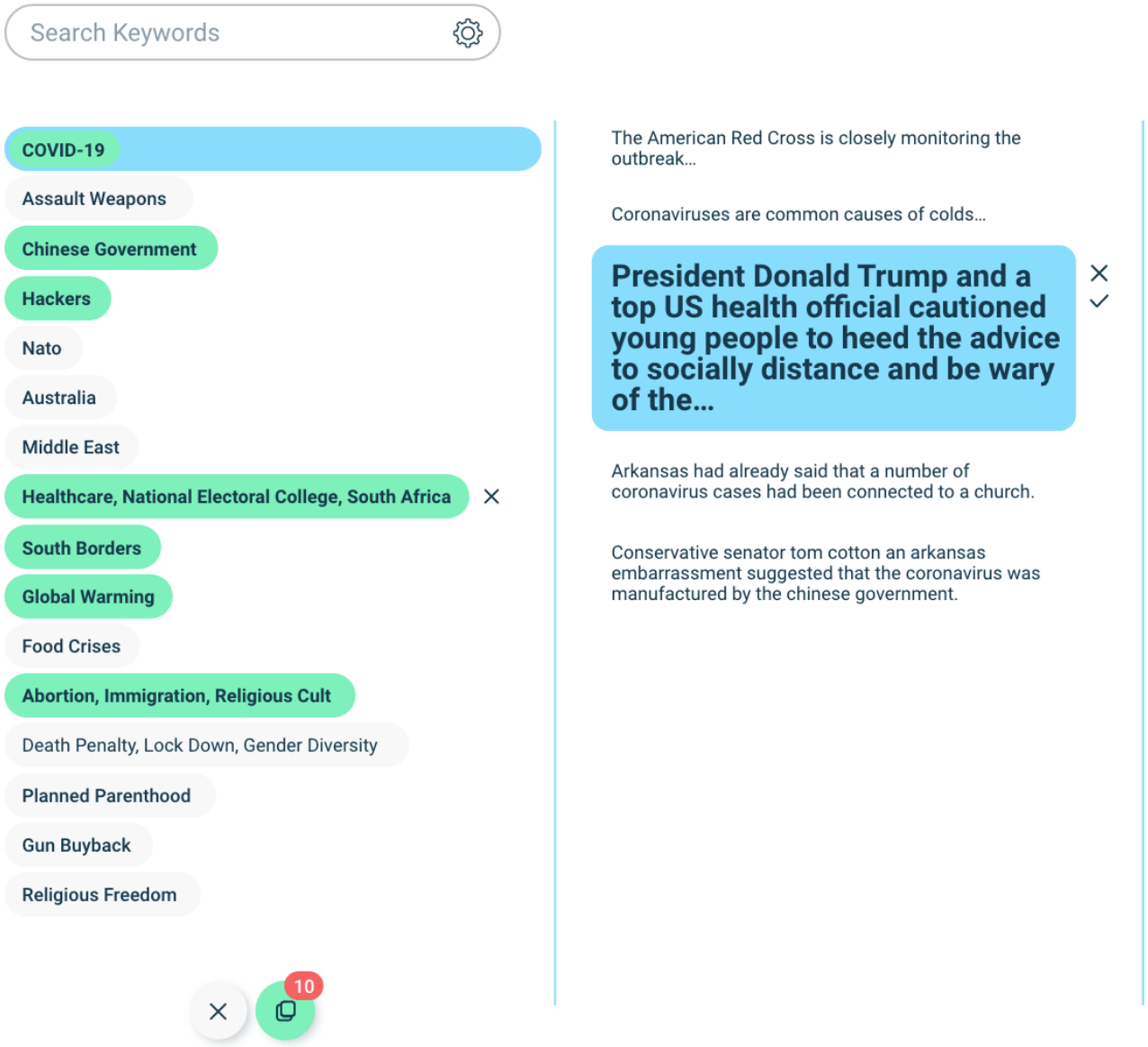

Figure 3: Narrative visualization tool listing top entities and their associated narratives. Compound narratives are shown with comma separated keywords

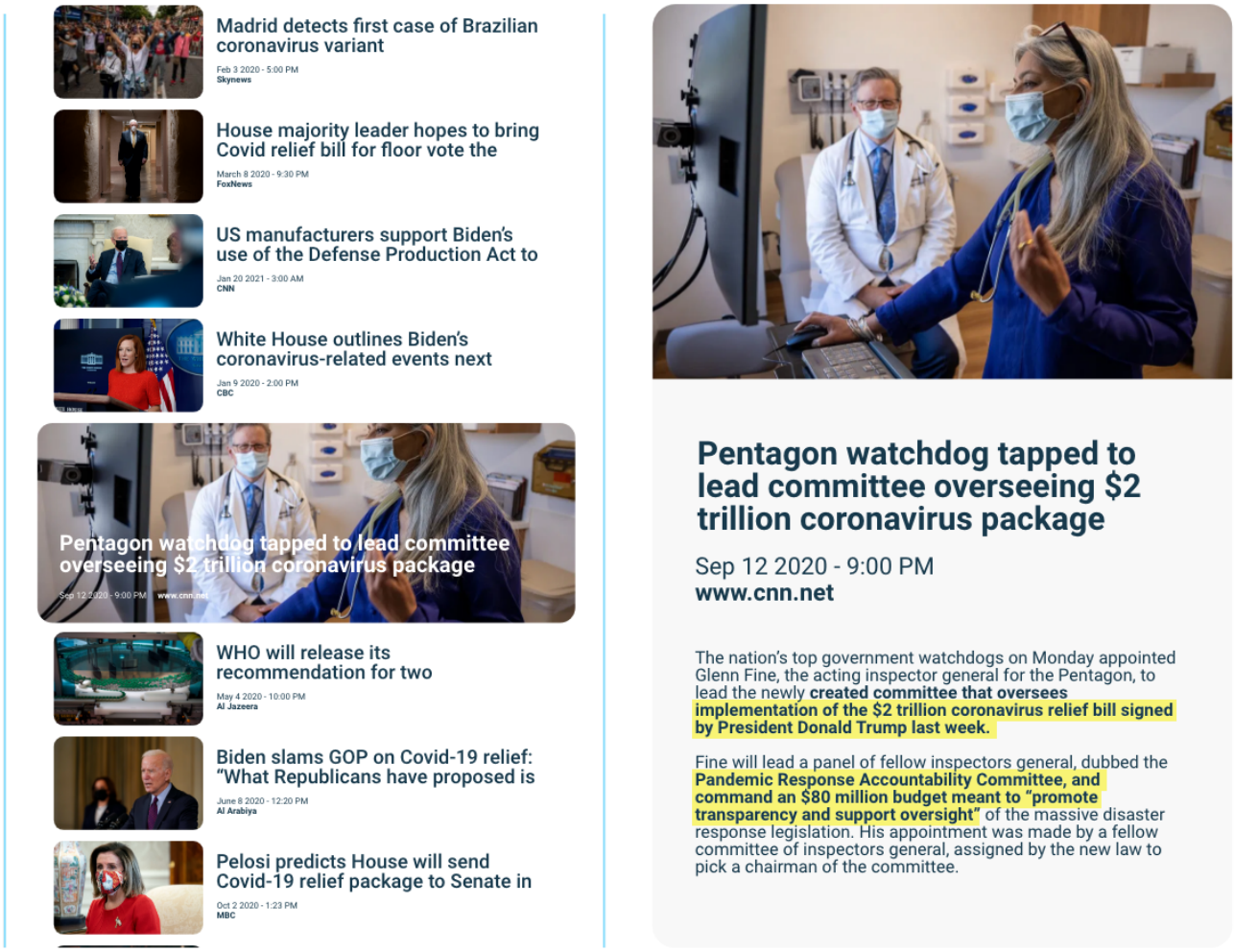

Figure 4: Narrative visualization tool listing the blog posts associated with the narrative

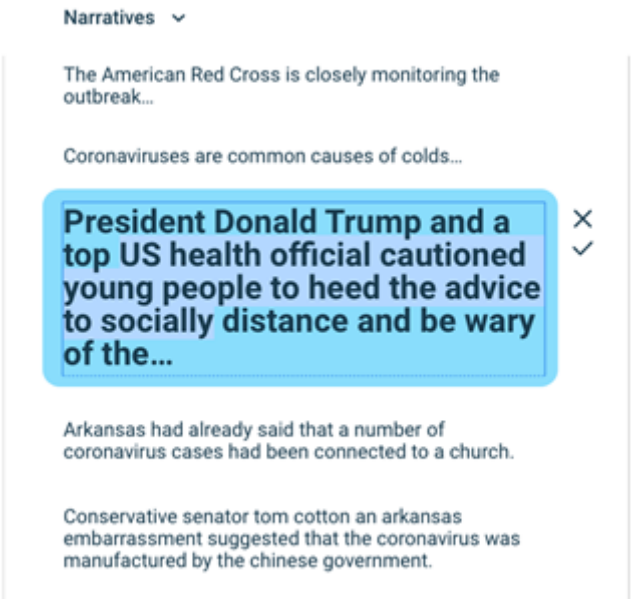

Figure 5: Narrative visualization tool showing capability for a user to edit a narrative

We are advancing our narrative extraction algorithm by making it faster, more accurate, and platform independent. Furthermore, we intend to add several features including ability to track evolution of narrative. Being able to observe multiple narratives change over time could help identify their origins as well as prominent events after which narratives evolve by merging, splitting or even completely flipping, as observed in our previous study [HBAkA18]. This will also help identify periods where a fringe narrative becomes dominant and vice versa.

Acknowledgement

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-17-S-0002, W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

[BHA20] Kiran Kumar Bandeli, Muhammad Nihal Hussain, and Nitin Agarwal. A Framework towards Computational Narrative Analysis on Blogs. In *Text2Story@ ECIR*, pages 63–69, 2020.

[Blo18] Blogtrackers. https://btracker.host.ualr.edu/, 2018.

[CHFA11] Wen-Ying Sylvia Chou, Yvonne Hunt, Anna Folkers, and Erik Augustson. Cancer survivorship in the age of YouTube and social media: a narrative analysis. *Journal of medical Internet research*, 13(1), 2011.

[CRF12] S. Corman, Scott W. Ruston, and Megan Fisk. A pragmatic framework for studying extremists' use of cultural narrative. In *2nd International Conference on Cross-Cultural Decision Making: Focus*, volume 2012, pages 21–25, 2012.

[DVK19] Anne Dirkson, Suzan Verberne, and Wessel Kraaij. Narrative Detection in Online Patient Communities. In *Text2Story@ ECIR*, pages 21–28, 2019.

[HBAkA18] Muhammad Nihal Hussain, Kiran Kumar Bandeli, Samer Al-khateeb, and Nitin Agarwal. Analyzing Shift in Narratives Regarding Migrants in Europe via Blogosphere. In *Text2Story@ ECIR*, pages 33–40, 2018.

[KNM20] Laura Koivunen-Niemi and Masood Masoodian. Visualizing narrative patterns in online news media. *Multimedia Tools and Applications*, 79(1):919–946, 2020. Publisher: Springer.

[MA20] Simone Mellace and Alessandro Antonucci. Temporal embeddings and transformer models for narrative text understanding. In *Text2Story@ ECIR*, 2020.

[MHRL+17] Sean McKenna, N Henry Riche, Bongshin Lee, Jeremy Boy, and Miriah Meyer. Visual narrative flow: Exploring factors shaping data visualization story reading experiences. In *Computer Graphics Forum*, volume 36, pages 377–387. Wiley Online Library, 2017. Issue: 3.

[Min18] Nat Gyenes Mina, An Xiao. How Misinfodemics Spread Disease, August 2018. Section: Technology.

[Mir] Holmstrom Miranda. Miranda Holmstrom. The narrative and social media. | StratCom.

[RCS+16] Scott W. Ruston, J. V. Cohn, S. Schatz, H. Freeman, and D. J. Y. Combs. More than just a story: Narrative insights into comprehension ideology and decision-making. *Modeling Sociocultural Influences on Decision Making: Understanding Conflict, Enabling Stability*, page 27, 2016.

[SML18] Shane Sheehan, Masood Masoodian, and Saturnino Luz. Comfre: a visualization for comparing word frequencies in linguistic tasks. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, pages 1–5, 2018.