Visualization Viewpoints

Editor:
Theresa-Marie Rhyne

Topic- and Time-Oriented Visual Text Analysis

Wenwen Dou
University of North Carolina at Charlotte

Shixia Liu
Tsinghua University

From manufacturing to education, retail, service, healthcare, and government, text analysis and understanding plays a crucial role to further growth, productivity, and innovation and thus is at the center of economic activities. To facilitate the process of converting textual data into actionable knowledge, visual text analysis has become a popular topic with active research efforts contributed by researchers worldwide. Numerous text visualization systems have been developed for analyzing various sources of textual data, including social media, news articles, scientific publications, patents, and Wikipedia.

In this article, we first present the benefit of combing text analysis (topic models in particular) with interactive visualization. We then highlight examples from prior work on topic- and time-oriented visual text analysis. Lastly, we present four challenges that warrant additional future research in the field of visual text analysis.

Advantages of Visual Text Analysis

The lack of structure and the high noise ratio in textual data makes their analysis particularly challenging. To address this, researchers have developed many text mining algorithms to summarize and analyze large amounts of textual information.[1] Although automated text-analysis methods are extremely powerful, it is still advantageous to tightly couple the analysis algorithms with interactive visual interfaces for two reasons. First, the output from the automated analysis is often too complex for data analysts to consume. Second, the sense and decision making based on the topic results have to rely on end users, with the tasks often being exploratory and iterative. Therefore, by integrating text analysis methods with interactive visualizations, we can empower users to explore, analyze, and make sense of the trends and sometimes unexpected patterns that are otherwise buried in massive amounts of texts.

Topic-Driven Visual Text Analysis

Topic models are designed for discovering the main themes that pervade a large and unstructured text collection.[1] Since the inception of topic models, a number of research efforts have been devoted to producing topics that both capture the characteristics of the corpora and are easy to interpret.[1] Each topic consists of a group of terms that, when combined together, synthesizes a higher-level meaning. The output of the topic models usually involves sets of keywords that form topics and multiple probabilistic distributions describing the rich relationship among topics, documents, and keywords.

Several visual text analysis systems leverage the power of topic models. Visual analysis of topic results can help address many common tasks: Which topics/themes summarize the text corpora? What are the relationships between the topics/themes? How do we find documents that focus on a specific topic of interest?

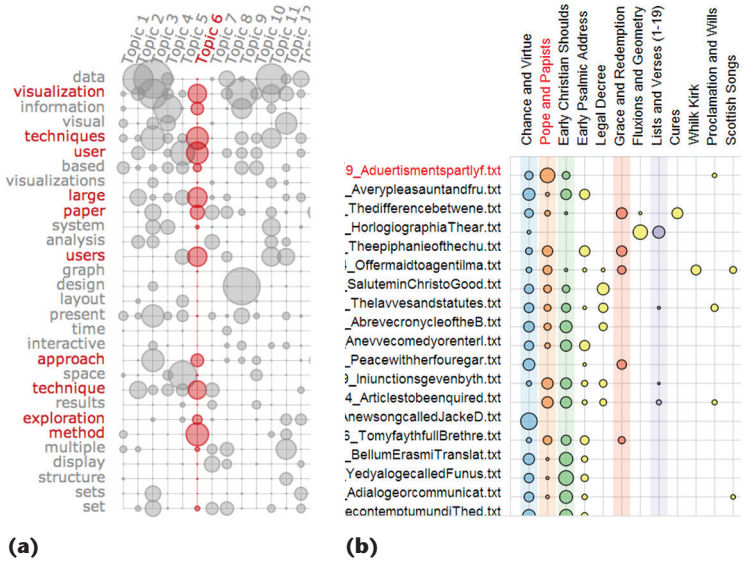
Visualizations that have been adapted and applied to revealing the document-topic-term relationship include parallel coordinates and matrices. Figures 1a and 1b show matrix visualizations that illustrate the correspondences between topics and terms[2] and between topics and documents,[3] respectively. The matrix visualization enables easy identification of large proportions based on circle sizes and reranking of the terms or documents based on proportions.

In addition to matrix visualizations, researchers have also adapted parallel coordinates to represent topic proportions (see Figure 2).[4] Brushing on the

Published by the IEEE Computer Society

high range on an axis/topic selects documents with high proportions on the topic. Such interaction enables filtering down to a set of documents that are highly relevant to the topic of interest.

As the text corpora grows (to billions of tweets, millions of research publications, and so on), the number of topics to capture the themes that pervade the text collections will increase accordingly. Therefore, it could be challenging to examine and navigate through a topic space, even though it is already an order of magnitude smaller than the original document space. To alleviate the problem of topic abundance, researchers have organized a flat list of topics into a hierarchical structure based on the similarity of topics.[5–7] Users can then collapse, expand, or modify the topic hierarchy. This hierarchical topic organization enables users to navigate the topic space and reorganize the topic hierarchy based on their mental models.



(a) (b)

Figure 1. Matrix visualizations. The sizes of the circles represent the proportion of the correspondence between (a) topics and terms and (b) topics and documents. (Left image courtesy of Jason Chuang and his colleagues,[2] and right image courtesy of Eric Alexander and his colleagues[3])

Time-Oriented Visual Topic Analysis

Topics capture themes that pervade a text collection. Because each document is often associated with a time stamp, the topics distilled from the documents exhibit different volumes over time. Visualizing topic trends is one of the many benefits of combining interactive visualization with topic models. Such visualizations support multiple tasks. For example, how do the topic trends evolve over time? Are there emerging topics? Can events be identified based on topic trends? How do topics merge and split? Is there any topical lead-lag effect between different corpora?

Many methods utilize a river or stack graph metaphor to portray evolving topics over time. For example, TIARA employs an enhanced stacked graph to illustrate how topics evolve over time (see Figure 3).[8] Similarly, the theme river metaphor[4]

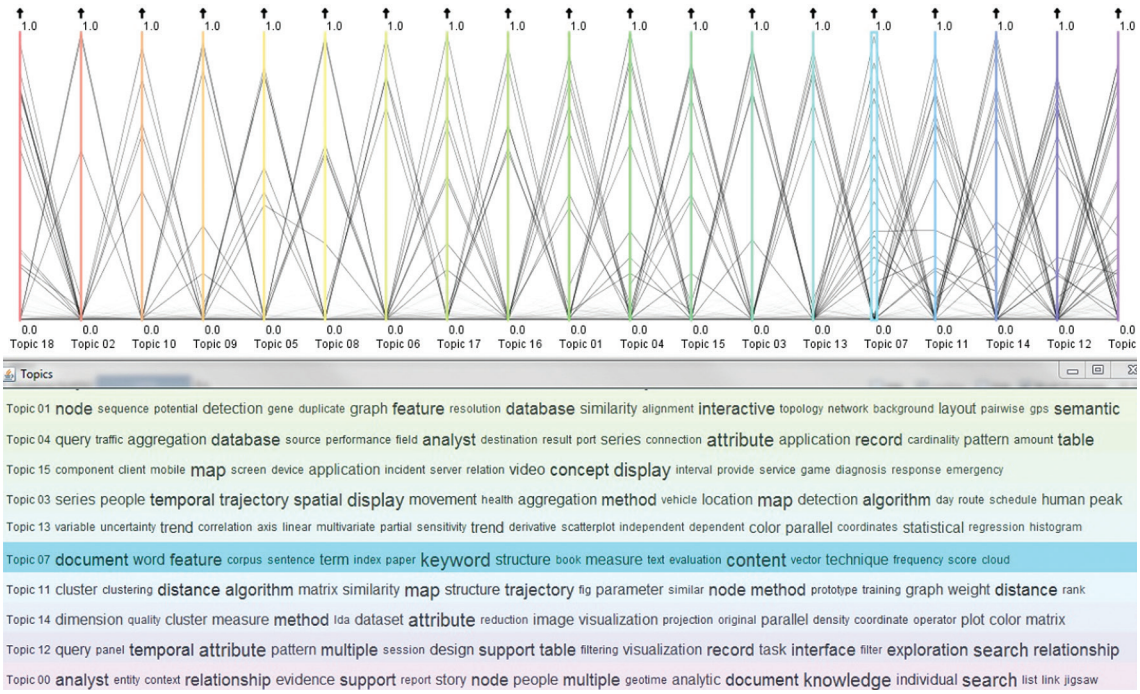


Figure 2. Adapting parallel coordinates to visualizing topic proportions. Each axis represents a topic, with the content of the topic shown in the bottom view.

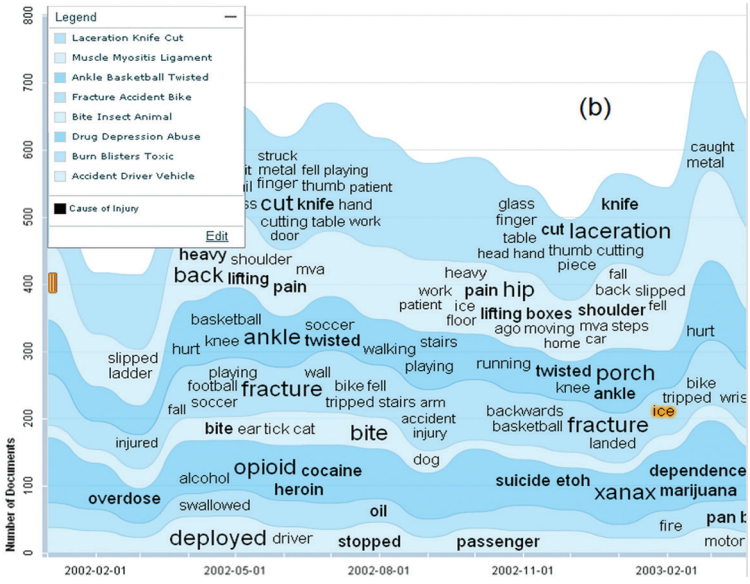


Figure 3. A stacked graph metaphor for visualizing topical trends. Using an enhanced stacked graph, TIARA illustrates how the topics evolve over time.[8]

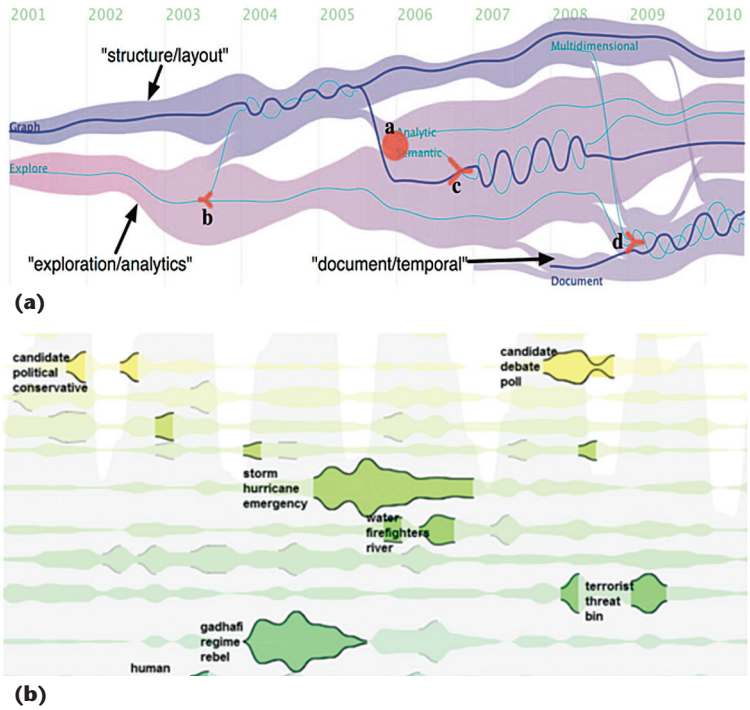


Figure 4. Visualizing relationships in textual data. These visualizations (a) explore topic splitting and merging and (b) represent events that are discovered from topic streams.

represents topic trends over time. In these visualizations, each ribbon represents one topic, so users can interactively explore individual and overall topical trends.

In addition to visualizing topic trends over time, other visualization systems look at how the relationships among different topics evolve. TextFlow was developed to help analysts visually analyze topic merging and splitting relationships over time (Figure 4a).[9] This metaphor also enables the identification of critical events that might have led to the merging or splitting of topics. In addition to splitting and merging, evolving topics may be correlated in other ways. Such correlations include topic lead-lag relationships, competition, and collaboration. In many other applications, it is desirable to identify which text corpus (lead) is followed by others (lags) regarding a specific topic. The visual analysis tool TextPioneer tightly integrates interactive visualization with lead-lag analysis to help users better understand lead-lag relationships across corpora both globally and locally.[10]

In addition to visualizing topics as continuous trends, researchers have performed further extraction based on the topic trends and identified discrete events. Events can serve as a succinct summary of large text corpora and are often identified by detecting bursts from topic trends. The Leadline visual analysis system also enables users to explore and analyze events that are described in text collections.[11] Figure 4b highlights individual events that usually occur in a much shorter time period than the overall time range of the corpus. Topic keywords are reranked to identify words that most accurately describe the specific events. As a result, users can examine and analyze the events that are computationally extracted from the text collection.

Challenges and a Call for Action

The field of visual text analysis is still young, and many interesting topics are still worth exploring. Figure 5a provides a pipeline that most of the aforementioned visual analysis systems follow, although some of the arrows haven't been realized yet. For example, the interaction arrow pointing to the text-mining component has yet to be fully explored (see the next section for a more detailed discussion). In Figure 5b, textual data are categorized by the "variety" property. Most of the aforementioned visual text analysis tools focus on homogeneous textual data. The exceptions include the lead-lag visual analysis system[10] that leverages metadata in addition to pure texts and the work on streaming text analysis.[12]

Interactively Modifying Topic Modeling Results

One of the advantages of combining interactive visualization with text-analysis techniques algorithms is the possibility of incorporating user feedback to improve the algorithms. In the case of visual analysis based on topic modeling, user feed-

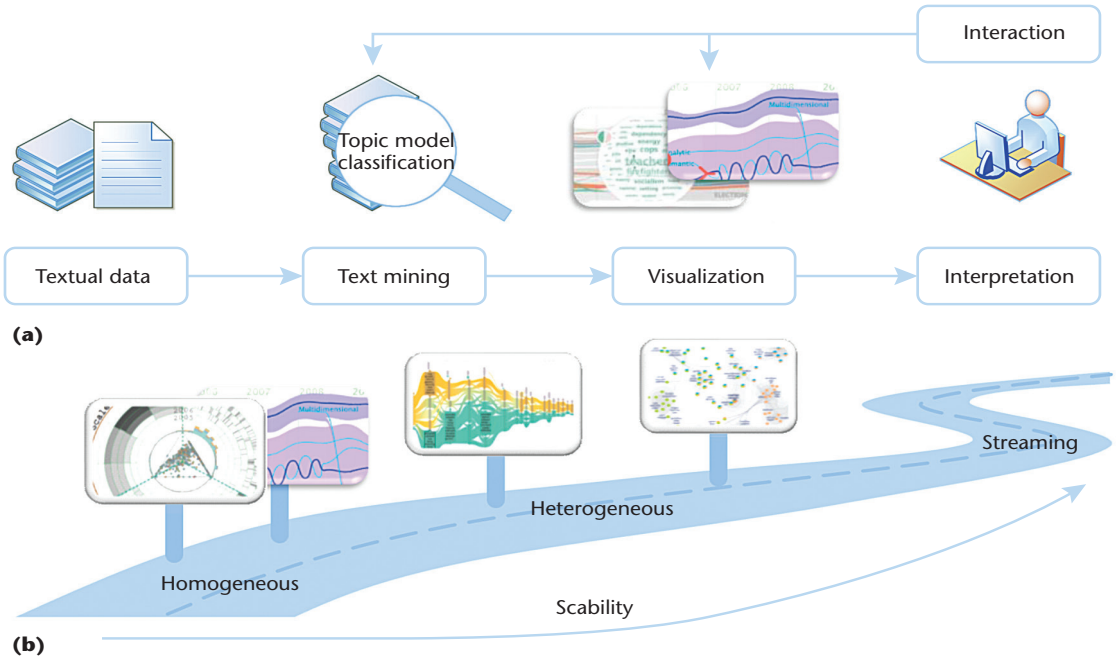


Figure 5. Visual text analysis systems. (a) A common pipeline for visual text analysis systems. (b) Visual text analysis systems that address different types of textual data.

back can be especially valuable on the semantic coherence of the topics. So far, few research studies have attempted to leverage user input through interactive visualization to improve the topical results. The main reason is that probabilistic topic models are computationally intensive. Although online versions of the topic models are available that do not require a retraining every time new documents come in, it is still unlikely that they can produce topic results in real time in order to meet the needs of interactive visualization.

One possibility is to leverage user feedback via topic-based visual analysis systems to modify the intermediate results as opposed to retraining the model. The feedback could be on the relationships between keywords in topics (for example, showing that certain keywords do not make sense to appear together) as well as the relationships between document and topics. The intermediate results that could be modified based on user input include the multinomial distributions of keywords and topics. A measure needs to be put in place that, if the resulting modified probabilistic distribution is different enough from the original distribution, such input needs to be considered when retraining the model at a later time.

Another challenge in interactively providing feedback to topic models via visual interfaces is determining how to translate user interactions (such as moving two keywords together or moving a keyword from one topic to another) into constraints that modify the probabilistic distribution. In addition, how does changing one multinomial distribution impact the rest of the probabilistic distributions?

Visual Analysis of Streaming Textual Data

Some textual data arrives in constant streams. Such data include social media posts, news articles, and emails. Exploring and finding patterns in streaming data in real time can help users discover emerging patterns. The streaming property not only requires analytical methods that can analyze incoming data, it also calls for proper visual metaphors that present how new data fits into the current trends and patterns. Shixia Liu and her colleagues presented an online visual analysis system for text streams. Their system integrates a streaming tree cut algorithm and a new visual representation that supports interactive analysis of streaming texts. The visualization integrates a sedimentation-based metaphor into the river flow metaphor to illustrate how new documents are aggregated into old documents (Figure 6).[12]

More work is still needed in the area of analyzing and visualizing streaming textual data. For example, when much of the new data is accumulated, we need to determine how to best update the current visualization that may have separated new data from historical data.[12] On the analysis side, the topic trends should be mostly kept consistent but still allow for major alterations if changes are discovered in the new data. On the visualization side, the update should be gradual so that users don't feel overwhelmed by multiple patterns all changing at once.

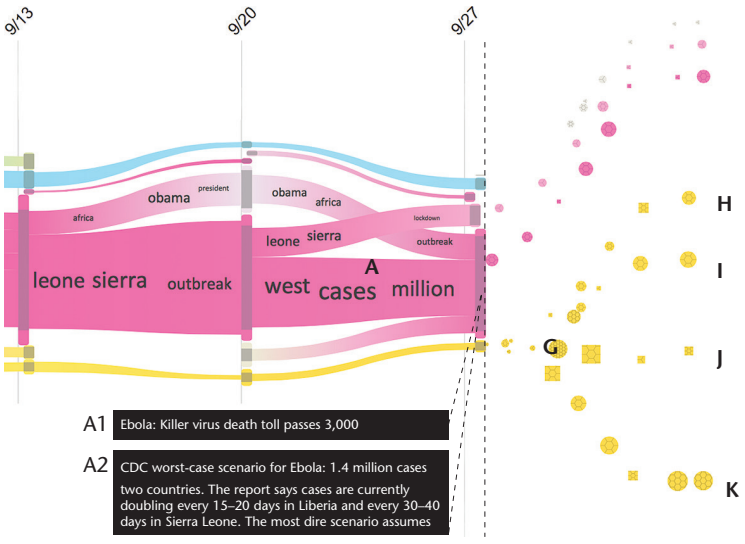


Figure 6. Visual analysis of streaming textual data. Integrating river flow and sedimentation visual metaphors allows us to illustrate how new documents are aggregated into old documents.[12]

Analysis of Heterogeneous Text-Based Data

In many applications, users often need to analyze heterogeneous textual data from multiple sources and in varying formats. For example, given a large amount of customer feedback from online review sites, social media, and emails, a product manager may want to know several things: Are new product features well received by customers? What are the major deficiencies of the current release? Which functions should be improved compared with the competitors' products? Product managers rely on such results to improve their products in preparation for subsequent releases. Analyzing textual data alone can't solve these problems, however. A taxonomy that captures the product features and functions should be analyzed together with textual data so that the results are more relevant to the questions asked.

In addition to textual data, documents often contain images or videos, which are also important for users to understand the entire text corpora. Accordingly, how do we tightly integrate textual data, images, and videos to form a full picture of an event/topic? The hypothesis is that jointly modeling texts and images/videos would bootstrap the understanding of each data type. The challenge of analyzing heterogeneous data together lies in identifying a common feature space that can represent various data types.

Scalability

Large volumes of data always present challenges for analysis, visualization, and interaction, and textual data is no exception. Large amounts of textual data require longer to analyze and call for scalable visual metaphors or interaction techniques to present an overview versus details. The online versions of topic models and the dynamic tree-cutting algorithms proposed in earlier work[12] are attempts to speed up the processing of large text collections by analyzing new data and its relationship with older data, as opposed to analyzing the entire collection all at once. On the visualization side, visualization researchers have utilized hierarchical topic structures to make the visual representation of topics scalable.[5–7] More research is needed to develop new scalable visual metaphors or adapt existing visualization techniques to present large text collections based on topics. ThemeRiver, for example, was adapted to show temporal trends of topics groups as an overview.[5] Details on individual topic trends are revealed on demand through user interactions. As the text collections grow large, the number of topics needed to capture the themes that pervade the collections also increases. More research is necessary to develop visual metaphors that can help end users navigate through the topic spaces for analysis.

The visual text analysis approaches discussed here focus on leveraging topic models, which assume a bag-of-words (BoW) model of texts. Compared with syntax-based approaches, the BoW model does not account for the order and sentence structure.

Visual text analysis has become a popular topic over the recent years, and as the challenging future research areas mentioned here are addressed, we may soon reach a point that methodologies and visual metaphors proposed for visual text analysis need to be evaluated more quantitatively. Thus, measures and benchmark datasets would greatly help the research community evaluate, categorize, and compare visual text analysis systems.

References

1. D.M. Blei, "Probabilistic Topic Models," *Comm. ACM*, vol. 55, no 4, 2012, pp. 77–84.
2. J. Chuang, C.D. Manning, and J. Heer, "Termite: Visualization Techniques for Assessing Textual Topic Models," *Proc. Int'l Working Conf. Advanced Visual Interfaces* (AVI), 2012, pp. 74–77.
3. E. Alexander et al., "Serendip: Topic Model-Driven Visual Exploration of Text Corpora," *Proc. IEEE Conf. Visual Analytics Science and Technology* (VAST), 2014, pp. 173–182.
4. W. Dou et al., "ParallelTopics: A Probabilistic Approach to Exploring Document Collections," *Proc.*

IEEE Conf. Visual Analytics Science and Technology (VAST), 2011, pp. 231–240.
5. W. Dou et al., "HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies," *IEEE Trans. Visualization and Computer Graphics*, vol. 19, no. 12, 2013, pp. 2002–2011.
6. W. Cui et al., "How Hierarchical Topics Evolve in Large Text Corpora," *IEEE Trans. Visualization and Computer Graphics*, vol. 20, no. 12, 2014, pp. 2281–2290.
7. S. Liu et al., "TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis," *ACM Trans. Intelligent Systems and Technology*, vol. 3, no. 2, 2012, article no. 25.
8. F. Wei et al., "TIARA: A Visual Exploratory Text Analytic System," *Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2010, pp. 153–162.
9. W. Cui et al., "Textflow: Towards Better Understanding of Evolving Topics in Text," *IEEE Trans. Visualization and Computer Graphics*, vol. 17, no. 12, 2011, pp. 2412–2421.
10. S. Liu et al., "Exploring Topical Lead-Lag Across Corpora," *IEEE Trans. Knowledge and Data Eng.*, vol. 27, no. 1, 2015, pp. 115–129.
11. W. Dou et al., "LeadLine: Interactive Visual Analysis of Text Data through Event Identification and Exploration," *Proc. IEEE Conf. Visual Analytics Science and Technology* (VAST), 2012, pp. 93–102.
12. S. Liu et al., "Online Visual Analytics of Text Streams," *IEEE Trans. Visualization and Computer Graphics*, preprint, 2015, doi:10.1109/TVCG.2015.2509990.

Wenwen Dou *is an assistant professor in the Department of Computer Science at the University of North Carolina at Charlotte. Contact her at wdou1@uncc.edu.*

Shixia Liu *is an associate professor in the School of Software at Tsinghua University. Contact her at shixia@tsinghua.edu.cn.*

Contact department editor Theresa-Marie Rhyne at theresamarierhyne@gmail.com.