A Survey on Existing Visualizations for Text and a Short Reflection on Semantics

Lanz Railey A. Fermin
Department of Information Systems and Computer Science
Ateneo de Manila University
Quezon City, Philippines
lanz.fermin@student.ateneo.edu

I. Introduction

Technological advancements throughout the last half-century has democratized vast quantities of information to nearly everyone, if not all, provided that they have a device connected to the Internet. Much of this accessible information is continuously growing, with no signs of slowing down, as [1] reported that at least 0.4 zettabytes (i.e., 4×10^{20} bytes) of data are created everyday. They typically appear in free-form, disorganized, and unstructured formats ranging from simple numeric tables to multi-modal content such as audio, video, image, and most notably, text that comprises approximately 80% of the total data volume ever created [2]. This has opened new and unimagined doors that allowed scholars to track emerging patterns in human thought and affect, shifts in ideology, and even the evolution of language use. To make sense of these large swathes of textual information, there is a critical need for tools that can automatically process and extract meaning from unstructured text; for it would be very inefficient, or impossible even, to rely on manual techniques for analysis.

Systematic methodologies for text analysis and associated procedures first appeared in the disciplines of social sciences, focusing on qualitative assessments and interpretation of existing documents [2]. However, the rise of the Internet and the rapid upswing on the computing powers of machines prompted a paradigm shift towards digital analysis, which was deemed "practical and inevitable" given the exponential increase of available data. Unlike numerical data, the digitization of text analysis proved difficult for one particular reason, among others: a computer cannot readily grasp contexts surrounding the word nor even create connections among words strung together in an input sentence.

Driven by developments in machine learning, the discipline of natural language processing (NLP) has emerged and quickly rooted itself as one of the premier fields at the intersection of artificial intelligence, information technology, and linguistics. As defined by [3], NLP is an area of research concerned with "processing natural languages ... and translating into numbers that a computer can use to learn about the world". True enough to this definition, NLP has made remarkable strides in recent years, enabling machines to perform a wide array of language-related tasks: ranging from basic sentiment analysis and text summarization, up to generating human-like responses in conversations with large language models and chatbots. These capabilities are made possible by transforming raw text into structured representations that machines can interpret and learn from. With the sheer volume and unstructured nature of textual data, there exists a need to transform it not only to a form that a machine can understand, but also to a form that humans can ingest to extract insights.

In this short paper, we survey some of the existing visualization techniques employed hand-in-hand with NLP such as word clouds, topic models, and networks. Moreover, we also note the limitations of current visualization technologies in conveying deeper semantic structures and contextual interconnections among words.

II. Data Visualization

Data visualization refers to the transformation of raw, complex, and sometimes unstructured, data into graphical representations like graphs, charts, and maps [4]. This makes information more accessible and interpretable to the viewer, as abstractions latent in the dataset are converted into tangible and concrete ideas. Hence, trends, patterns, and anomalies that are otherwise difficult to make sense of in purely textual data become easier to detect and predict.

Meanwhile, in certain pipelines or workflows, data visualization also helps in supporting faster and better-informed decision-making especially in time-sensitive situations. Given its capability to concretize abstractions, large datasets can be effectively highlighted through relevant correlations, groupings, and outliers. This way, the said process realizes two purposes: explanatory and exploratory, as it clarifies known results while also facilitating the discovery of new insights.

These functions of data visualization are crucial in NLP. The unstructured, and oftentimes messy, nature of textual data poses unique challenges because of its semantic richness and contextual variability, especially when dealing with languages composed of numerous polysemous words like Tagalog. In this field, data visualization techniques are applied to help understand word frequency, co-occurrence, topic distribution, among others. By translating language into statistics, then into visuals, researchers can better understand how certain words are distributed across documents or corpora.

There exists a specialized area within NLP known as visual text analytics, which primarily focuses on leveraging static and dynamic visualizations to support knowledge discovery from text. It is seated at the intersection of human cognition and computational analyses, relying on the graphic's capacity to perceive meaningful patterns and spatial relationships in abstract data. Visual text analytics emphasizes not only the display of structured attributes but also the depiction of semantic content; thus, offering new ways to explore large textual collections beyond the traditional and linear reading.

Given the importance of visualization in managing and interpreting text data, the next section surveys some of the most commonly used visualization techniques in NLP and visual text analytics, assessing their strengths and limitations.

III. Commonly-used Data Visualizations for Text

A. Word Clouds

A word cloud, sometimes referred to as a tag cloud, is defined as a visual presentation of a set of words, in which various aesthetic properties such as color, size, and weight are used to represent features associated with a specific word [5]. Figure 1 shows an example of a word cloud generated via the `wordcloud` library in Python.

Fig. 1. A word cloud generated by using the Wikipedia NLP article as the dataset. Photo by Sigli Mumuni, 2021.

From the figure, quick insights can already be generated such as the dominance of the term "natural language" and its abbreviation, "NLP", throughout the article. It is this very reason why word clouds are popular exploratory visualizations for text data: they are easy to generate and interpret, even for non-technical viewers. Given that they efficiently show the most frequent words in a document or corpus, prominent themes can be easily inferred which, in turn, opens up to a slightly deeper analysis of the text. Additionally, word clouds are language-agnostic; meaning, it can be generated regardless of the text's language provided that a frequency count can be done to it. This flexibility can be even extended to multilingual texts.

Despite its ease-of-use and insight-ready nature when observed, word clouds falter due to its lack of contextual presentation of the most frequent words. Before a word cloud is generated, the document is tokenized and treated as a word in isolation. With the existence of words that carry multiple functions and/or meaning within a document, it offers no semantic understanding, no consideration of sentence-level dependencies, and can lead to potential misinterpretation.

B. Latent Topic Visualizations

Using the Latent Dirichlet Allocation (LDA), a generative probabilistic model, large collections of documents can be analyzed by modeling them as a mixture of multiple topics; which, in turn, is represented as a distribution over a fixed vocabulary of terms [6]. This methodology allows for the unsupervised discovery of latent themes within textual data. By examining patterns of word occurrence across this large collection of documents, a latent topic modeling method is hereby created.

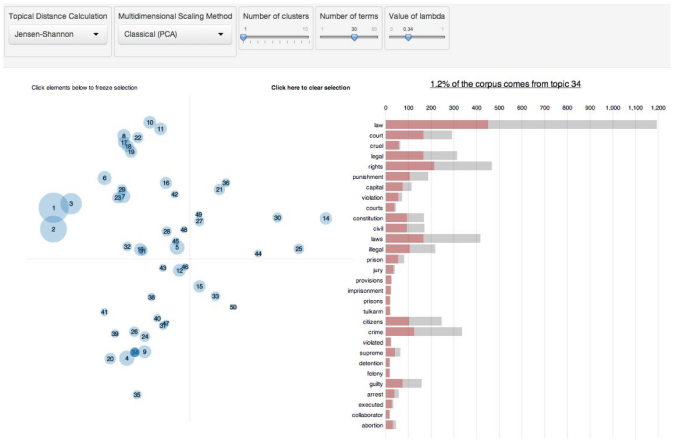

Fig. 2. A layout of the visual generated through the Python library `LDAvis`. Topical distances, scaling methods, and topic-term relationships are shown. Photo by Carson Sievert and Kenneth Shirley, 2014.

Topic visualizations, such as the one shown in Figure 2, provide a compelling kind of insight as they offer visual relational mapping between inferred topics. In `LDAvis`, a two-pane visualization is generated: a projection of topics into the 2D space using a multidimensional scale, and a depiction of salient terms for a selected topic shown via a bar chart. This enables a relational illustration of how topics are distributed across the dataset and what terms comprise the topics inferred by the model; hence, allowing a simultaneous qualitative and quantitative understanding of the text.

However, LDA-mediated visualizations tend to only show surface-level semantics, since proximity of words is derived from a statistical distribution rather than a linguistic context. As a result, discursive or truly thematic nuances are left vulnerable for ignorance. Furthermore, the two-pane visualization may also be cognitively overwhelming due to information overload, especially when too many topics cluster together or overlap. And finally, LDA models may suffer from overfitting as the number of parameters in the model grows linearly with the size of the corpus [7].

C. Graph-based and Network Visualizations

Network graphs are slowly rising in popularity as methods for visualizing relationships between named entities within

a document, or for depicting linguistic connections between words called collocations. These visualizations draw from concepts of graph theory, using nodes to represent entities and edges to depict the relationships between them. Figure 3 shows a collocation network depicting a homograph pair in Tagalog, and their shared words within a Tagalog news article collection dataset.

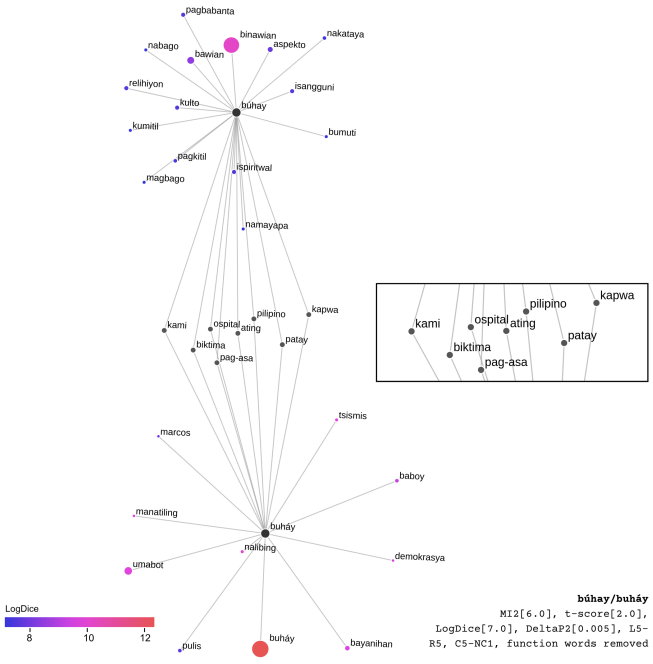

Fig. 3. A collocation network for the Tagalog homograph pair *búhay* (life) and *buháy* (alive). Photo by Lanz Fermin, 2025.

The presentation of textual data using network graphs are particularly helpful in exploratory analysis and/or delineation of conceptual linkages among words. From their constructions based on simple text statistics, like co-occurrence matrices and association measures for collocations, these networks are easy to generate and can be made interactive to support filtering by term frequency or zooming into specific sub-networks, allowing users to focus on salient themes. Their visual intuitiveness also makes them accessible for a broad audience, especially when applied to thematic mapping in large document sets.

That said, network graphs are also fallible to certain limitations. Like word clouds, its dependence on frequency often fail to capture the deeper semantic context in which terms occur. For collocations and the use case of homograph pair relationship shown above, it may be sufficient; but observing for how the word actually functions within the sentence, particularly for those homographs that may have multiple meaning within each variant, this type of visualization may not be sufficient. Moreover, networks also lack representation of syntax or directionality, ignoring word order and grammatical relationships that are crucial in understanding textual data.

IV. On Visualizing Semantics

As we have seen in the previous section, a recurring limitation observed in the contemporary visualization techniques is their inability to sufficiently capture the contextual and semantic richness of language. In their methods of transforming text into visuals, important elements like contexts, nuances, and other discursive markers are oftentimes easily lost due to the inherent frequentist approach employed by their respective algorithms. This results in a flattening of language, where the dynamic nature of words are lost and meanings become overly simplified. In fact, even the more advanced models like the LDA approach fall to the same trap: the bag-of-words assumption within its framework pushes it to treat documents as unordered sets of words.

All hope is not lost, however, as there are budding innovations that may provide a better view of contexts and semantics for NLP. The sudden uptick of attention directed towards transformer-based models have opened promising avenues toward the creation of a more context-aware visualization for text data. This can be attributed to their capability to harness large volumes of data with fairly viable efficiency through contextual, dynamic word embedding that map words into high-dimensional semantic spaces [8]. These spaces, though, are projected back again into 2D or 3D using dimensionality reduction techniques, which circles us back to the earlier problem present in the current technologies.

These, nevertheless, are encouraging signs but I posit that there may be some hurdles that should be addressed first, before we can fully actualize a dynamic and context-aware visualization for text data. The first hurdle is that visualization tools must be able to natively integrate dynamic embeddings or transformer-based outputs that adapt word meaning to specific contexts. Second, new methods for dimensionality reduction and layout generation must be developed to preserve relational integrity while maintaining clarity and interpretability of words and their resulting text statistics. Lastly, the sheer complexity of language demands interactive, multi-layered visualizations that can readily adapt to various drill-down approaches: based on time, genre, or user-defined granularity.

References

[1] K. Bartley, "Data Statistics (2025) - How much data is there in the world?," Rivery. Accessed: Jul. 23, 2025. [Online]. Available: https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world/

[2] R. Egger and E. Gokce, "Natural Language Processing (NLP): An Introduction," in Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications, R. Egger, Ed., Cham: Springer International Publishing, 2022, pp. 307–334. doi: 10.1007/978-3-030-88389-8_15.

[3] H. Lane and M. Dyshel, Natural Language Processing in Action, Second Edition. Simon and Schuster, 2025.

[4] N. Cao and W. Cui, "Overview of Text Visualization Techniques," in Introduction to Text Visualization, N. Cao and W. Cui, Eds., Paris: Atlantis Press, 2016, pp. 11–40. doi: 10.2991/978-94-6239-186-4_2.

[5] M.-T. Chi, S.-S. Lin, S.-Y. Chen, C.-H. Lin, and T.-Y. Lee, "Morphable Word Clouds for Time-Varying Text Data Visualization," IEEE Transactions on Visualization and Computer Graphics, vol. 21, no. 12, pp. 1415–1426, Dec. 2015, doi: 10.1109/TVCG.2015.2440241.

[6] C. Sievert and K. Shirley, “LDAvis: A method for visualizing and interpreting topics,” in Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, J. Chuang, S. Green, M. Hearst, J. Heer, and P. Koehn, Eds., Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 63–70. doi: 10.3115/v1/W14-3110.

[7] D. M. Blei, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[8] N. Patwardhan, S. Marrone, and C. Sansone, “Transformers in the Real World: A Survey on NLP Applications,” Information, vol. 14, no. 4, Art. no. 4, Apr. 2023, doi: 10.3390/info14040242.