2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Understanding Massive Text Visualization

1st Giovana Voltoline
Computer Science Departament
University of Beira Interior
Covilhã, Portugal
gi.voltoline@ubi.pt

2nd Sebastião Pais
University of Beira Interior and NOVA LINCS
Covilhã, Portugal
sebastiao@di.ubi.pt

3rd Bruno Silva
Instituto de Telecomunicações
University of Beira Interior
Covilha, Portugal
bsilva@di.ubi.pt

4th João Cordeiro
University of Beira Interior
INESC TEC Porto
Covilhã, Portugal
jpcc@ubi.pt

***Abstract*—Massive text visualization is a burgeoning field that addresses the visualization, exploration, and analysis of extensive textual datasets, encompassing various domains such as social media, scientific literature, news articles, and more. This paper overviews recent advances, challenges, and potential solutions in massive text visualization and systematically categorizes these works, focusing on shared characteristics such as objectives, preprocessing techniques, processing approaches, and visualization methods. This comprehensive analysis can better understand the current and emerging trends in massive text visualization.**

***Index Terms*—Massive Text, Text Visualization, Textual Data**

I. Introduction

The Digital 2023 Global Overview Report, produced by DataReportal in partnership with Meltwater and We Are Social, revealed that 64.4% of the worldwide population uses the internet and that the average amount of time spent daily on it is 6 hours and 37 minutes. The same report also shows that the two main reasons people use the internet are finding information and staying in touch with friends and family. Regarding the latter, roughly 60% of the world's population uses social media. Moreover, Twitter, for example, has a 280-character limit for most of its users, and about 350,000 tweets are sent every minute. This is just a small fraction of the constantly produced textual data.

So, in the digital age, an overwhelming amount of textual data is generated daily from various sources, such as social media, and text visualization emerges as a powerful approach to navigating through the immense sea of textual data, transforming it into visual representations humans can comprehend, interpret, and leverage for informed decision-making. Massive text visualization is a multidisciplinary field that combines elements of natural language processing (NLP), data visualization, and machine learning to extract meaningful patterns, trends, and relationships from unstructured text.

This paper aims to find what has been done in the area of Massive Text Visualization. To achieve this, we conducted a comprehensive and systematic review of papers that covered the theme of Big Textual Data Visualization. Our primary focus was on papers dealing with textual data that incorporated some form of data visualization to ensure the inclusion of essential concepts.

II. Massive text visualization

Massive text visualization is a dynamic and evolving field that focuses on representing and exploring large-scale textual data. With the ever-growing volume of digital content, such as social media posts, news articles, scientific publications, and online documents, the need to derive meaningful insights from this vast sea of unstructured information has become increasingly critical [1].

At its core, massive text visualization aims to transform complex textual data into visual representations easily interpretable by humans. By leveraging a combination of natural language processing, data visualization techniques, and machine learning algorithms, massive text visualization enables researchers, analysts, and organizations to discover patterns and correlations hidden within the vast amount of text.

One of the primary challenges in massive text visualization is the sheer volume of data. Traditional text analysis methods are ill-equipped to handle such massive datasets efficiently. Therefore, scalable and parallel processing algorithms are crucial to enable real-time exploration and interaction with the data [2].

The unstructured nature of text also presents a significant obstacle. Unlike structured data, which fits neatly into rows and columns, text lacks a standardized format. It contains multiple languages, idioms, slang, and context-specific jargon, making it challenging to organize and represent coherently. NLP techniques, including tokenization, part-of-speech tagging, and named entity recognition, are essential for transforming raw text into structured data.

Another critical aspect of massive text visualization is abstraction and summarization. As the volume of data grows, presenting all the details becomes impractical and overwhelming. Therefore, advanced machine learning algorithms, such as topic modelling and text clustering, help identify key themes

This work has been partially funded by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT.IP and the European Commission through the Horizon 2020 project Pharaon, grant agreement no. 857188.

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
https://doi.org/10.1145/3625007.3627482

and topics within the text, creating concise and informative visual summaries.

Interactive visualization is vital in empowering users to explore massive text datasets effectively. Interactive interfaces allow users to filter, drill down, and zoom in on specific aspects of the data, revealing deeper insights and supporting data-driven decision-making. Moreover, combining text with other data modalities, such as images, time series data, or geographical information, enriches the visualization and provides a more comprehensive understanding of the underlying patterns and connections.

Despite the numerous advancements and promising applications, challenges remain. Ensuring data privacy and ethical considerations when dealing with sensitive textual information is paramount. As text data often contains personally identifiable information, researchers and practitioners must adopt privacy-preserving techniques to safeguard users' identities.

In conclusion, massive text visualization is a powerful and indispensable tool for making sense of the vast amounts of textual data generated daily. By integrating cutting-edge technologies, such as NLP, machine learning, and interactive visualization, this field enables us to derive valuable insights, discover trends, and gain a deeper understanding of the complex world of text. As technology continues to evolve, so will the capabilities of massive text visualization, revolutionizing how we analyze and interact with textual information.

III. Related Work

The primary goal of this paper is to categorize the selected papers into distinct categories, such as objectives, preprocessing, processing approach and data visualization. This section thoroughly summarises some selected papers to facilitate this understanding, offering critical insights into their respective contents. We used primarily practical papers to enhance the reader's comprehension of the concepts explored in this survey. This approach also allows us to introduce important concepts that will be further examined in subsequent sections, providing a foundation for a more in-depth analysis.

A. Tweetviz - Twitter Data Visualization

Published in 2014 by Stojanovski et al. [3], the purpose of Tweetviz is to enable users to visualize Twitter data effectively. Tweetviz offers a range of interactive visualizations, allowing users to gain insights into various aspects, such as Twitter users, keywords, and hashtags. By utilizing the Tweetviz interface, users can input their desired data and explore different visualization techniques. The tool's visualizations intend to provide valuable information and aid in understanding patterns and trends within Twitter data.

The authors acknowledge the growing popularity of social media platforms and highlight the increasing prevalence of microblogging services. Specifically focusing on the Twitter community, they note its wide range of uses in daily life, including sharing thoughts, discussing popular topics, disseminating news, and even marketing.

Given the vast volume of daily tweets generated on Twitter, the authors are motivated to extract valuable knowledge regarding user interests and behaviour. They aim to detect trends and patterns in information dissemination within this data source. Recognizing the immense potential for insights hidden within this dataset, the authors aim to uncover meaningful information that can contribute to a deeper understanding of user behaviours and preferences.

To develop Tweetviz, several components were needed: a user interface, visualization techniques, and topic distribution analysis. The authors utilized a Python framework for implementing the Latent Dirichlet Allocation (LDA) topic distribution. In terms of visualization, they leveraged tools such as Google Charts and d3.js. The authors opted for a combination of HTML, CSS, and JavaScript for the front interface.

Acquiring the data is the initial step in data processing; the Twitter REST API was utilized for this purpose. Although the authors acknowledged that the API does not index all tweets, it still provided a substantial number of tweets sufficient for the project.

In conclusion, while this paper may not extensively delve into the processing steps, it offers a comprehensive showcase of five different visualizations. These visualizations include distributions, word clouds, and streamgraphs for the LDA topic distribution. Through these visual representations, the paper successfully presents the data processing and analysis outcomes, shedding light on the underlying patterns and insights derived from the dataset.

B. Twitter Data Clustering and Visualization

In 2016, Sechelea et al. [4] introduced a comprehensive system that collects geotagged tweets, performs data processing and analysis, and visualizes the outcomes. The incorporation of geotagged tweets explicitly facilitates the creation of a three-dimensional (3D) environment for the visualization of identified trends [4]. The paper emphasizes that geotagging represents an invaluable opportunity to verify trends and patterns within distinct geographical locations.

For data collection, the authors used the Twitter REST API and a Python script to filter and gather geotagged tweets. The analysis phase employed Python, MapReduce, and Hadoop. The visualization aspect leveraged Gephi to depict topic proximity and OpenStreetMaps for mapping. The authors also employed a clustering method to identify dominant themes within the dataset that combined the k-means algorithm with the Density-Based Spatial Clustering of Applications with Noise (DSCAN) algorithm.

In terms of visualization, the paper presents seven distinct visualizations, including 2D/3D maps and clusters. The authors conclude by emphasizing that this system serves as a prototype and lays the foundation for future work across various domains.

C. Twitter Data Visualization and Sentiment Analysis

The work of Patil et al. [5] aims to leverage Twitter to verify people's opinions regarding revoking Article 370 of

the Indian Constitution. Additionally, the authors explore its impact on trade in Pakistan, terrorism in Pakistan and India and the opinions on Kashmir related to Article 370. To achieve this, they collected tweets related to these themes and categorized them as positive, neutral, or negative, providing insights into public sentiment [5].

When writing this paper, Twitter had already amassed 100 million users and generated a remarkable 500 million tweets daily. Given the vast amount of data available and the techniques of opinion mining, the authors recognized this as an opportune moment to investigate and comprehend the impact of the revocation of Article 370 on both India and Pakistan. By harnessing the power of Twitter data analysis, the authors aimed to gain valuable insights into the public sentiment and reactions surrounding this event.

For data collection, the authors utilized Tweepy to gather tweets from August 5th to August 30th. Sentiment analysis was conducted using TextBlob, a Python library. In terms of visualization, the authors employed matplotlib and pandas, both Python libraries, to effectively present their findings.

This paper offers a comprehensive analysis through six distinct visualization techniques, including bar graphs, line graphs, histograms, and a pie chart. The authors' analysis uncovers meaningful patterns and insights, enabling them to draw informed conclusions regarding the investigated theme. Based on their analysis, the authors highlight that Pakistani users displayed more concern regarding the impact of Article 370's revocation on trade, whereas Indian users expressed heightened apprehension about the escalation of terrorism. However, it is noteworthy that despite these concerns, the authors found that Indian users generally maintained a positive perspective on the revocation.

D. Supervised Sentiment Analysis of Twitter Handle of President Trump with Data Visualization Technique

In 2020, Kalyan et Al. [6] conducted a study to explore the potential influence of then-president Donald Trump's social media usage on the public approval of his administration. The researchers analyzed tweets posted by the president over approximately three years and compared them with approval ratings obtained from two different websites. By examining this data, the authors aimed to ascertain whether there existed a correlation between the president's social media activity and the public's approval. The Twitter data utilized in this project was collected from January 2016 to January 2019 through an archive website that organizes postings based on date, time, and device. The approval rate data was obtained from The Real Clear Politics and cross-checked with other reliable sources.

This paper presents seven diverse visualizations, which enable the authors to draw several conclusions. Firstly, they observe that the use of Twitter does not immediately impact approval ratings; instead, a lag of approximately five days is observed when comparing datasets. Furthermore, the authors suggest that while this approach may be less practical for assessing mass political opinion than consumer products, it could yield more detailed insights when combined with additional data sources such as newspapers and visual media reports.

E. Tweetviz - Visualization Tweets for Business Intelligence

Sijtsma et al. [7] introduce Tweetviz, an interactive tool designed to extract valuable business information, offering visualizations of sentiment analysis on tweets related to business locations, identify other businesses frequented by customers, and estimate customer profiles.

The authors collected vast data from multiple datasets, comprising 24 million geotagged tweets from the San Francisco Bay Area. This data spanned from June 2013 to March 2015 and was stored in a MySQL server. To estimate customer addresses, the authors leveraged Zillow and Flickr. The system's front end was developed using the Angular JavaScript framework, while Node.js supported the back end.

This paper includes a screen capture showcasing the Tweetviz interface, which allows users to visualize a map displaying various business venues and a sentiment scale ranging from negative to positive.

F. Analysis and Visualization of Twitter Data Using k-means Clustering

Garg et al. [8] focused on the behaviour and structure of a social network by utilizing geotagged tweets. The study involved extracting, refining, analyzing, and visualizing these tweets in a geospatial representation. The purpose of incorporating geotagged tweets was to enable the visualization of tweet clusters in specific geographical locations, providing insights based on contextual information.

This study employed OAuth to extract the required data, initially obtained in JSON format and converted into a data frame format. The analysis, clustering, and visualization tasks were performed using sixteen distinct R packages.

Six distinct clusters were identified across India after the processing, extending slightly beyond its borders. The visualizations offered in the study comprised three map-based visualizations and a word cloud.

G. Factors Driving the Popularity and Virality of COVID-19 Vaccine Discourse on Twitter - Text Mining and Data Visualization Study

In 2021, Zhang et al. [9] investigate the popularity and virality of tweets related to the COVID-19 vaccine using text-mining techniques, and also examine the topic communities of the most liked and retweeted tweets through network analysis and visualization [9]. The authors emphasize that the topic generated extensive discussions following the declaration of the COVID-19 outbreak as a pandemic by the World Health Organization in March 2020. The approval of the initial two vaccines further fueled debates, including concerns regarding vaccine safety and effectiveness. Moreover, the authors noted the impact of political polarization in the United States, reinforcing the arguments. Given the abundance of data available on social media platforms, particularly Twitter, the authors were motivated to delve into the topic using this rich source of information.

The data collection for this study focused on tweets from US-based users and specifically included English-language content related to the COVID-19 vaccination. The dataset consisted of a total of 501,531 tweets, which were collected between January 1, 2020, and April 30, 2021.

For this study, the authors utilized topic modeling to identify latent topics within the tweets. They also conducted sentiment analysis to classify the general sentiment expressed in the tweets. In addition, for the 2500 most liked tweets and 2500 most retweeted tweets, the authors employed network analysis and visualization techniques to detect the topics and explore the relationships between them. The study provided two different network visualizations.

H. A Machine Learning Approach for Disease Surveillance and Visulization using Twitter Data

Ashok et al. [10] explore the potential of Twitter as a tool for detecting real-time events, with a focus on disease surveillance. The study emphasized the significance of such surveillance in empowering the public to adopt preventive measures, aiding pharmaceutical manufacturers in enhancing sales of disease-specific medications, and ensuring their timely availability. Health organizations can leverage the findings to implement necessary measures for controlling outbreaks.

This study gathered a comprehensive collection of tweets between January 2018 and June 2018, utilizing the Twitter REST API, focusing on Dengue, Norovirus, H1N1, Zika, Influenza, and Ebola. Subsequently, the acquired data was subjected to cleaning procedures using StreamListener and stored in a JSON file. To facilitate further analysis, preprocessing techniques were applied to the data. Finally, clustering techniques were employed to process the tweets.

In addition, this paper presented two visualizations to enhance understanding. Firstly, a distribution map was created to illustrate the geographical spread of the various diseases. Secondly, a heatmap was generated, depicting the intensity of these diseases worldwide and visually representing their prevalence.

IV. Approaches

Given that our analysis consisted of the papers, it is natural to expect some divergences in the researchers' objectives for their respective datasets. Thus, the purpose of this section is to delineate the main objectives observed in the literature and provide an example of each objective using a selected paper. Despite significant variations in the objectives among the papers, it was still possible to identify common themes and group them into distinct categories, including:

a) Trend Finding: A simple technique used to identify the most common words or phrases in a piece of textual data. It involves analysing the frequency of specific words or phrases within the text and ranking them in order of importance. This technique helps identify a text's main ideas or themes and better understand the content. This paper [3] presents an analysis of Twitter data, including keywords and hashtags, as well as user-specific information such as posting frequency, most commonly used hashtags, and frequently used words.

b) Topic Modelling: It uses statistical algorithms to group words based on their co-occurrence patterns in the text and then clustering these groups into meaningful topics. In paper [9], a Latent Dirichlet Allocation model was used to identify the leading topics in a data set of 501,531 tweets related to the COVID-19 vaccine. The analysis resulted in the identification of 12 major topics.

c) Sentiment Analysis: It can be employed, for instance, to understand customers' feedback without needing to read every individual comment manually. The objective of [6] is to investigate whether there is a correlation between President Donald Trump's tweets and his approval rating over three years. To achieve this, the authors classified each president's tweets as negative or positive on a scale ranging from -1.0 to 1.0.

d) Geo Location: With the ability to track device location on social media platforms, many papers aimed to visualize the trending topics in certain regions. In paper [10], the authors aimed to investigate the feasibility of tracking diseases using Twitter by analysing tweets that mentioned diseases over five months. In addition, they utilised geotagged tweets to determine the location of the diseases.

e) Business Intelligence: Many business owners are interested in understanding their customers' opinions on the products they are selling and their overall customer profile. Textual data from social media, emails, or product feedback can provide valuable insights to help uncover this information. The paper [7] aims to provide business owners with insights related to their business, such as the sentiment towards specific venues' locations and their customers' profiles.

V. Processing the Data

As demonstrated earlier in Section IV, papers often exhibit significant variations in their objectives. However, many of these papers share similar preprocessing and processing approaches. Therefore, this section encompasses the most commonly encountered preprocessing and processing methods that we have identified.

a) Preprocessing: Texts may include uninformative words, like articles and prepositions, that would be beneficial to filter out, among other preprocessing steps, which may include: *case normalization*, *irrelevant term removal*, *spelling correction*, *tokenization*, *stemming*, and *lemmatization*.

b) Processing Approach: While the objectives of the papers varied greatly, the processing techniques mainly focused on one of three processes. The most commonly used approaches included: topic modeling with *Latent Dirichlet Allocation* (LDA), different forms of *Sentiment Analysis*, and *Data Clustering* methods, mainly *K-Means*, *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), and *Agglomerative Clustering*.

In the literature, we came across various techniques, with some papers utilizing only one technique while others employing multiple techniques, particularly in the preprocessing stage. The tables below summarize the papers based on the Preprocessing Techniques and Processing Approach they used.

TABLE I
PREPROCESSING. EACH ROW DISPLAYS THE PAPERS THAT UTILIZED THE PREPROCESSING TECHNIQUE.

Preprocessing Technique	Papers
Case Normalization	[4]–[6], [8], [10]–[12]
Irrelevant Term Removal	[4]–[6], [8]–[12]
Spelling Corrections	[6]
Tokenization	[4]–[6], [8]–[10], [12]
Stemming	[4], [6]
Lemmatization	[6], [9], [12]

TABLE II
PROCESSING APPROACH. EACH ROW DISPLAYS THE PAPERS THAT UTILIZED THE PROCESSING APPROACHES.

Processing Approach	Papers
Latent Dirichlet Allocation (LDA)	[3], [9], [11]
Sentiment Analysis	[5]–[7], [9], [13]
Clustering Methods	[4], [8], [10]

VI. VISUALIZATION TECHNIQUES

In simple terms, data visualization is a graphical representation of information. While not all of the papers we selected included this component, many did; for some, it was the primary focus. The visualization techniques used in the literature were generally similar.

a) Word Cloud: One popular technique is word cloud visualization, which visually represents the frequency of words in a given text collection. Word clouds provide a quick overview of the most frequently occurring terms, allowing users to identify prominent themes or topics within the text. By adjusting each word's font size or colour based on frequency, word clouds can highlight the most important terms, as shown in Figure 1.

b) Graphs: This method was frequently used in the literature (Figure 2), to represent various types of information

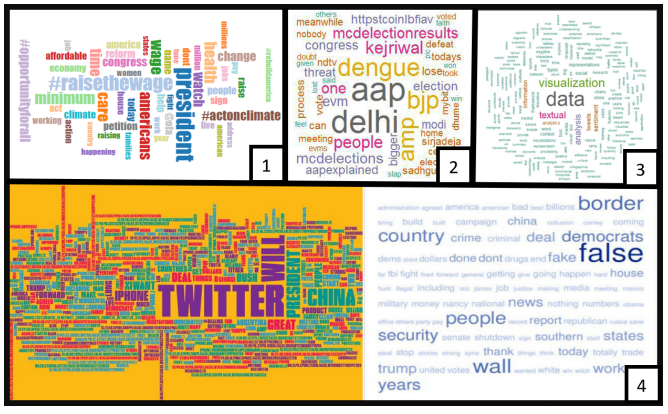

Fig. 1. Word clouds found in the literature. (1) Frequent keywords from paper [3], (2) Word frequency from [8], (3) Word cloud from paper [14], (4) Image to the left is word cloud before preprocessing and on the right after preprocessing from the paper [6].

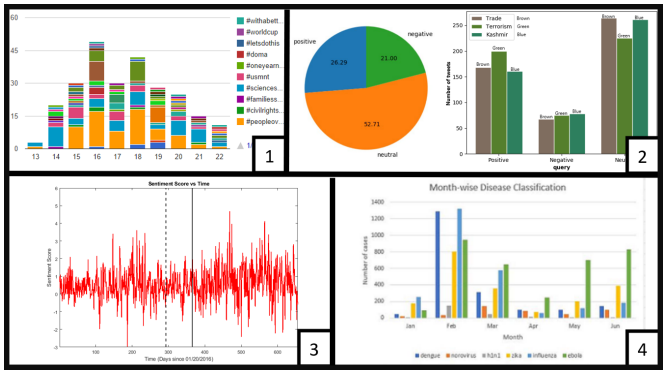

Fig. 2. Graphs found in the literature. (1) User-hashtag distribution from different times of the day [3], (2) Sentiment Analysis pie chart and Sentiment Analysis on topics from [5], (3) Sentiment score vs time [6], (4) Montwise disease classification and count [10].

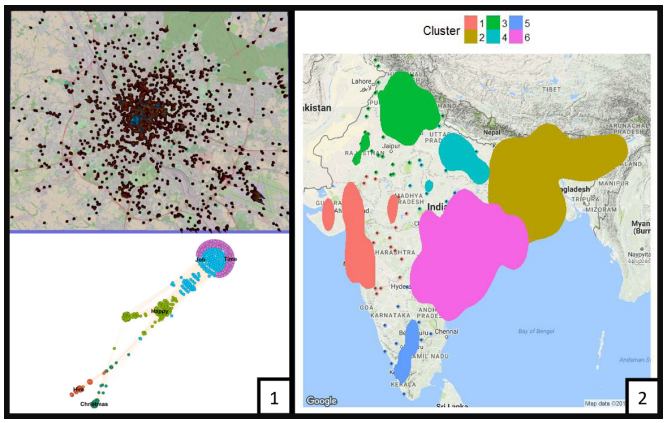

Fig. 3. Clusters visualization found in the literature. (1) Tweet Density for the city of Brussels and keywords given by clutering algorithm [4], (2) Overlay visualization of Clusters from the paper [5].

such as word frequency, topic distribution over time, and comparisons. It was commonly presented in the form of bar graphs, line graphs, or histograms.

c) Clusters: Various methods were used in the papers to graphically represent clusters, with the most frequent ones being bubble charts or visualizations within maps (Figure 3).

d) Others: Other visualization techniques were used less frequently in the papers, such as network diagrams, score plots, and distribution maps, as shown in Figure 4.

VII. CHALLENGES

Massive text visualization poses a significant challenge due to the vast amount of textual data that needs to be processed, analyzed, and presented meaningfully. Traditional visualization techniques may prove inadequate with millions, if not billions, of documents, articles, social media posts, and other textual sources available. Storing and processing such massive amounts of text require powerful computational resources and specialized algorithms capable of handling the complexity efficiently.

Another significant hurdle is the unstructured nature of textual data. Text lacks a predefined format or fixed vocab-

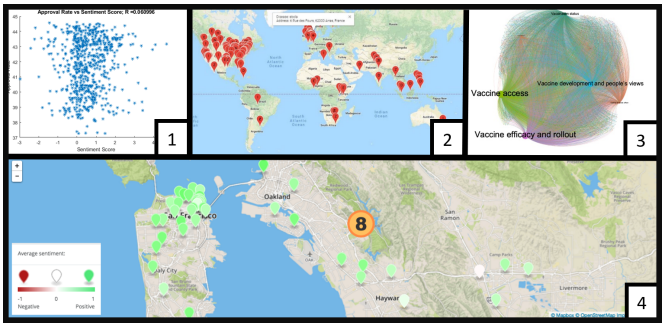

Fig. 4. Various types of data visualization found in the literature. (1) Approval Rate vs Sentiment Score from [6], (2) Map depicting the distribution of various diseases [10], (3) Topic communities of 2500 most retweeted tweets related to the COVID-19 vaccination [9], (4) Color corresponding to the location average sentiment [7].

ulary, making it difficult to represent and organize effectively. Natural language is nuanced, and capturing a visualization's subtle semantics, context, and underlying patterns demands sophisticated natural language processing (NLP) techniques.

Furthermore, ensuring the scalability and responsiveness of visualizations becomes crucial when dealing with massive text datasets. Users expect interactive and real-time data exploration, which can be a substantial technical challenge, particularly with limited resources and bandwidth.

The issue of information overload is also prominent in massive text visualization. Presenting too much information at once can overwhelm users, making it difficult to comprehend and draw insights from the data. Striking a balance between providing a comprehensive view of the information and avoiding cognitive overload requires careful design and thoughtful consideration of the visualization's layout and interactivity.

Moreover, addressing the problem of data quality and relevance is paramount. Large text datasets often include noise, redundant information, and irrelevant content.

Lastly, preserving data privacy and adhering to ethical considerations can be a significant challenge when working with massive text datasets. Sensitive information may inadvertently be exposed through visualizations, posing a risk to individuals' privacy and potentially leading to unintended consequences.

In conclusion, tackling the challenges of massive text visualization requires a multidisciplinary approach that combines advanced computational techniques, NLP algorithms, data cleaning methodologies, interactive visualization design, and ethical considerations.

VIII. Conclusion

In this paper, given the abundance of available data, we aimed to explore the existing research in massive text visualization. To achieve this, we comprehensively reviewed numerous papers, allowing us to understand the subject matter better. Through our analysis, we identified several common characteristics among the papers.

Most practical papers followed a consistent workflow, encompassing data acquisition, preprocessing, processing, and visualization. Building upon this observation, we categorized the papers based on these steps, excluding data acquisition while incorporating the objectives outlined in each paper.

By employing this categorization technique, we discovered that the most widely utilized processing approach was *Sentiment Analysis*. Furthermore, despite the significant variation in individual objectives, from analyzing public sentiment toward the COVID-19 vaccine to customer profiling, we could group these objectives into distinct categories.

The existence of such diverse objectives suggests that there are numerous possibilities for further exploration within this field. By leveraging the extensive range of textual data available, researchers can delve into topics such as topic modeling, trend finding, sentiment analysis, geo-location, and business intelligence.

References

[1] D. Angus, S. Rintel, and J. Wiles, "Making sense of big text: a visual-first approach for analysing text data using leximancer and discursis," *International Journal of Social Research Methodology*, vol. 16, no. 3, pp. 261–267, 2013.

[2] N. Cao and W. Cui, *Introduction to text visualization*. Springer, 2016, vol. 1.

[3] D. Stojanovski, I. Dimitrovski, and G. Madjarov, "Tweetviz: Twitter data visualization," *Proceedings of the data mining and data warehouses*, 2014.

[4] A. Sechelea, T. Do Huu, E. Zimos, and N. Deligiannis, "Twitter data clustering and visualization," in *2016 23rd international conference on telecommunications (ICT)*. IEEE, 2016, pp. 1–5.

[5] R. Patil, N. Gada, and K. Gala, "Twitter data visualization and sentiment analysis of article 370," in *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*. IEEE, 2019, pp. 1–4.

[6] K. Sahu, Y. Bai, and Y. Choi, "Supervised sentiment analysis of twitter handle of president trump with data visualization technique," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2020, pp. 0640–0646.

[7] B. Sijtsma, P. Qvarfordt, and F. Chen, "Tweetviz: Visualizing tweets for business intelligence," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 1153–1156.

[8] N. Garg and R. Rani, "Analysis and visualization of twitter data using k-means clustering," in *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2017, pp. 670–675.

[9] J. Zhang, Y. Wang, M. Shi, X. Wang *et al.*, "Factors driving the popularity and virality of covid-19 vaccine discourse on twitter: text mining and data visualization study," *JMIR public health and surveillance*, vol. 7, no. 12, p. e32814, 2021.

[10] A. Ashok, M. Guruprasad, C. Prakash, and S. Shylaja, "A machine learning approach for disease surveillance and visualization using twitter data," in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*. IEEE, 2019, pp. 1–6.

[11] Ž. Krstić, S. Seljan, and J. Zoroja, "Visualization of big data text analytics in financial industry: a case study of topic extraction for italian banks," *ENTRENOVA-ENTerprise REsearch InNOVAtion*, vol. 5, no. 1, pp. 35–43, 2019.

[12] K. C. K. Ven, A. N. K. Ying, N. Q. Jie, S. Y. Lun, S. L. C. Yuen, D. Handayani, N. Hamzah, M. Lubis, and T. Mantoro, "Depression identification through social media posts: Data preprocessing for data visualization of tweets," in *2021 IEEE 7th International Conference on Computing, Engineering and Design (ICCED)*. IEEE, 2021, pp. 1–6.

[13] A. Almjawel, S. Bayoumi, D. Alshehri, S. Alzahrani, and M. Alotaibi, "Sentiment analysis and visualization of amazon books' reviews," in *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*. IEEE, 2019, pp. 1–6.

[14] C. Conner, J. Samuel, A. Kretinin, Y. Samuel, and L. Nadeau, "A picture for the words! textual visualization in big data analytics," *arXiv preprint arXiv:2005.07849*, 2020.