**INF4000 Data Visualization Report**

**How does sentiment vary across different news categories, and does this suggest emotional manipulation in political news, especially during the election period?**

course code: INF4000

word count: 3233 words

reg number: 240175647

Composite Visualization

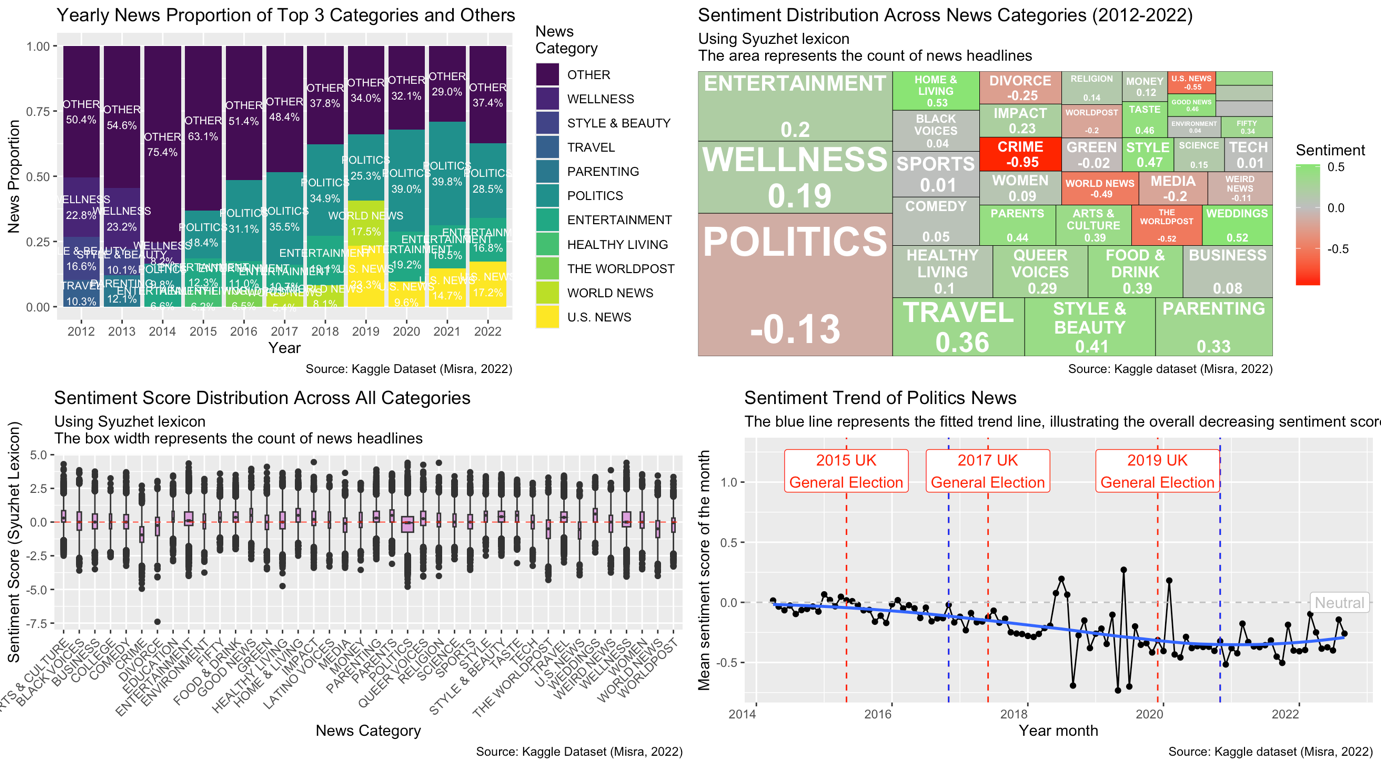
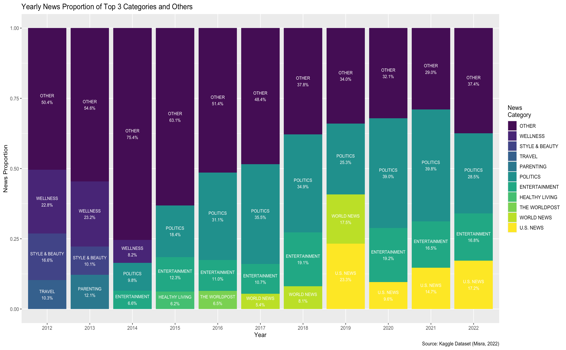


Figure 1 The composite visualization

Knowledge Building

News, according to Cambridge Dictionary, is the information presenting recent events with fair and objective views (Cambridge University Press, n.d.). It not only provides information to the public but also reflects, or even modifies, the public’s attitude towards, for example, political and economic activities (Evans & Lyons, 2008; Huang et al. 2010; Seifollahi & Shajari, 2019). In Evans and Lyons’s (2008) research, a 30% of daily variance in the stock market price is found out which is caused by macro-economic news. Research indicates that media can directly influence political opinions by exposing audiences to the content of specific broadcasts or newspaper articles (Norris el al., 2000; Hopmann et al., 2010). In political news, negative content is often employed to evoke public concerns, which can serve to push specific agendas and influence public opinions, such as shaping election choices. Furthermore, in sentiment analysis, many researchers achieve a precise rate of more than 80% when only studying the news headlines, instead of the whole passage of the news article (Huang et al. 2010; Peramunetilleke and Wong 2002; Nassirtoussi et al. 2015). To obtain a fair and sustainable news industry and avoid bias, it is important to study how news media use emotions in their news to attract readers, influence public perceptions and even shape public opinion. For the public, these insights could be beneficial in empowering them to evaluate news critically and provide them with knowledge on identifying potential biases or emotional manipulation in the news they consume. Also, researchers could be inspired by the ethical implications of emotional control in the industry.

The four visualizations presented above are composite representations of the findings from this sentiment analysis research, based on data from the News Category Dataset by Rishabh Misra (2022), retrieved from Kaggle. The first figure is the stacked bar chart of news categories across time, which shows the proportion of major news categories during the past decades. It is displayed that political, entertainment and U.S. news are the prominent ones in recent years. Then, a treemap and a boxplot provide information about the sentiment composition of and comparison between the most significant news categories. Most of the categories have a positive tendency on their headlines. In particular, home and living news is the most positive one, with a mean sentiment score of 0.53 and the majority of data points are distributed in positive areas, while crime news, with a score of -0.95 and the whole ‘box’ is under the neutral line, is one of the most negative news categories. Finally, the scatter plot showing the fluctuation of sentiment scores of political news between 2014 to 2022. It is clear that there is an overall descending trend of emotional tendency among political news and the news headline obtaining negative emotion most of the time.

Theoretical Frameworks

一張含有 文字, 行, 繪圖, 圖表 的圖片

自動產生的描述

Figure 2 Sentiment trend of politics news

An ASSERT framework is adopted to facilitate the development of visualizations that are clear, enlightening, informative, and engaging (Ferster, 2013). It includes six main steps: ask a question, search for information, structure the data, envision the answer, represent the visualization, and tell a story. This framework is used to build a scatter plot with line that shows the average sentiment score of news in the news category to answer the question: Is there a consistent or occasional emotional tendency in political news? The detailed description of ASSERT processes is stated in the Table 1 below:

|  |  |
| --- | --- |
| Ask | To ask a good research question, the prerequisite is to define the audience in advance (Ferster, 2013). As the aim of this study is to provide the public with the information about sentiment manipulation in political news, major considerations will be the needs and capacities of the audience and some other content-sensitive and accessible demands.  The primary target audience is the general public, who may seek a broad understanding of sentiment manipulation in political news and come from diverse backgrounds. Thus, the visualization will be kept simple, avoiding complex designs. For accessibility, considerations such as minimizing reliance on color will address issues like color blindness.  The question was developed using the six Ws (Sherry et al., 2021):   * Who: the public * What: the sentiment of political news * Where: in the UK * Why: to reveal the potential emotional manipulation in news * When: 2012 - 2022 * How: sentiment analysis of news title   Thus, the question is, “Is there a consistent or occasional emotional tendency in political news?”. |
| Search | To ensure good quality solutions to research questions, there are ways to acquire meaningful data (Kirk, 2012). A meaningful dataset should have sufficient volume and granularity with many independent variables and well-structured data (Few, 2009). The dataset used in this research is a secondary dataset from an open and publicly accessible database source in Kaggle (Rishabh Misra, 2022), which scrapes news data from HuffPost starting from 2012 and includes enough data and variables, such as date, headlines, short descriptions, and authors of 209,527 news articles from 2012 to 2022. This dataset is suitable as it provides sufficient temporal data for sentiment analysis, enabling a deeper exploration of patterns (Ferster, 2013). |
| Structure | In the structuring process, erroneous values are cleared, missing data is handled, and data formats are standardized to ensure consistency (Ferster, 2013). This process converts raw data into an analyzable format through cleaning and preprocessing. Data is typically categorized as:   * Continuous (e.g., sentiment scores), * Categorical (e.g., news categories for grouping).   Structured data must be error-free, complete, and well-organized for meaningful analysis. For example, categorical data requires proper coding. Since the news database has no missing values in the headline column, sentiment analysis is applied after text preprocessing, including stop words removal and lemmatisation. |
| Envision | Envision is a process of exploration in the early stages of design, including iterative comparisons, trends, and finding relationships, with the aim of gathering as many valuable ideas as possible without any critical judgement (Ferster, 2013).  In the present visualization, based on Edward Tufte’s principles of analytic design and statistics and informatics (Ferster, 2013), word frequency, sentiment analysis, and range, mean, and IQR value of sentiment scores are calculated to help analyze relationships, trends, and changes in this data. |
| Represent | The visualization was implemented using ggplot2 in R Studio. According to Norman’s action cycle (1988), the goal of this graph is to display the relationship of sentiments of political news and time. A simple scatter plot was selected as the visualization to represent the trend of the mean sentiment score of political news headlines from 2014 to 2022, as the pre-attentive perceptions and cognitive (short-term memory) limitation (Norman, 1988; Shneiderman & Plaisant, 2010) of humans. This type of plot is ideal for showing how average sentiment scores vary over time. The x-axis represents time (year-month), while the y-axis represents sentiment scores of political news. |
| Tell | The visualization follows Gustav’s narrative flow triangle, comprising the start, rising action, climax, falling action, and end (MacEwan, 1900). It begins with the title, “Sentiment Trend of Politics News,” inviting viewers to explore emotional trends. The rising action is established through data points and election reference lines, encouraging predictions about the relationship between elections and sentiment shifts. This builds to the climax, where the audience observes the overall downward trend and potential sentiment decline before election periods, fostering insights.  The falling action highlights the stabilization of sentiment trends post-election cycles and slight recovery after 2020. The conclusion ties the narrative together: “Overall, political sentiment has declined over time, especially during election periods, with signs of recovery after 2020.” The source annotation adds credibility to the presented story. |

Table 1 Detailed explanation of ASSERT framework

Accessibility

一張含有 文字, 螢幕擷取畫面, 收據, 設計 的圖片

自動產生的描述

Figure 3 Sentiment scores distribution

Accessibility in visualizations refers to the ability of ensuring the information of the figures or charts could be accurately received by all audiences, including those with special needs (Kim et al., 2021). It involves optimizing visual perception and cognitive abilities, focusing on elements like colour usage and chart simplicity to avoid confusion or misinterpretation (Kirk, 2012). It is crucial because it allows all users—including those who are visually impaired—to do complicated activities like reading, comprehending, and creating visualisations, particularly higher-order reading tasks (Kirk, 2019).

In order to attain accessibility, three crucial factors must be taken into account (Kirk, 2012). First, accessible content involves the design of diagrams and interactions. As Kirk (2012) explained, most standard visualizations, such as bar charts, line graphs, boxplot, and scatter plots, is famous, which accounting for around 80% usage, because of their straightforward design and widespread familiarity, making them easier for most users to interpret. Thus, a boxplot is adopted to show the distribution of sentiment scores of news categories as this graph is widely used and most people are familiar with it so there is no need to explain the chart format in detail.

Secondly, accessibility design should accommodate users’ cognitive styles and natural information processing, emphasizing visual perception and cognitive capabilities for greater inclusivity (Kim et al., 2021). In this visualization, plum-coloured boxes stand out against the white background for better visibility. The fill colour is purely decorative and does not convey data, so colour blindness was not a consideration. Titles and subtitles summarize the chart, while labelled x- and y-axes provide clarity. To prevent overlapping text, x-axis category labels are rotated 45 degrees, ensuring clear and easy reading of dense text.

The primary limitation of the boxplot is the overcrowding on the x-axis, which may hinder readability, particularly for users with visual impairments. According to Ware’s (2012) conceptual model of perceptual processing, visual memory is limited by active attention demands, allowing only a few objects to be effectively held at a time (Sampanes, Tseng & Bridgeman, 2018). One potential solution is to reduce the number of news categories displayed. However, to provide a comprehensive view and ensure users can access sentiment scores for any category of interest, all categories are included in the chart after well consideration.

To enhance clarity and help readers quickly locate the sentiment score distribution of their preferred news categories, the x-axis categories are arranged alphabetically rather than by median sentiment scores. This approach ensures that readers can easily find specific categories of interest while still observing the overall sentiment distribution and identifying outliers effectively. Additionally, a red dashed line at the neutral sentiment score (0) serves as a reference point, guiding viewers in interpreting the sentiment distribution across categories.

User orientation is another key accessibility factor, focusing on the level of information detail, sensory modalities, and assistive tools (Siirtola, 2019; Kim et al., 2021). While some users may require only a general overview (e.g., titles, axis labels, or legends), others might seek detailed insights. This boxplot accommodates varying needs by presenting both an overview, such as the median, and detailed information, including outliers. However, the visualization lacks additional guidance, which may make it challenging for non-professional viewers unfamiliar with boxplots to interpret the information. To address this, accessible design features have been incorporated, such as providing more context in the subtitle to aid interpretation.

Visualisation Choice

一張含有 文字, 螢幕擷取畫面, 鮮豔, 平行 的圖片

自動產生的描述

Figure 4 Yearly top 3 count news categories

Stacked bar charts are an effective visualization technique for displaying cumulative data attributes, helping viewers understand how individual values contribute to overall totals (Streit & Gehlenborg, 2014). Despite their simplicity, several key design considerations must be addressed to ensure their effectiveness (Munzner, 2016; Indratmo et al., 2018).

Firstly, stacked bar charts can become visually inefficient when displaying multiple attributes. As Indratmo et al. (2018) explain, an increase in attributes can result in visual clutter, making the chart difficult to interpret. Secondly, the use of colour plays a crucial role in the chart’s readability. Different colours are often assigned to bar segments to differentiate subcategories, but poorly chosen colours can hinder comprehension and exclude colour-blind viewers (Healey, 1996). Finally, stacked bar charts can pose challenges for comparing attributes. Users may need to analyse both the overall totals and the distribution of subcategories (Brehmer & Munzner, 2013). Achieving a balance between these two focuses is essential for effective visualization.

For example, as shown in Figure 2, the chart includes six subgroups in each column, leading to visual clutter. Additionally, the choice of colours does not account for colour-blind accessibility. This design makes it difficult for viewers to compare the total sales of different products or to analyse the distribution of sales channels within each product. Such shortcomings highlight the importance of careful design in making stacked bar charts effective and accessible.

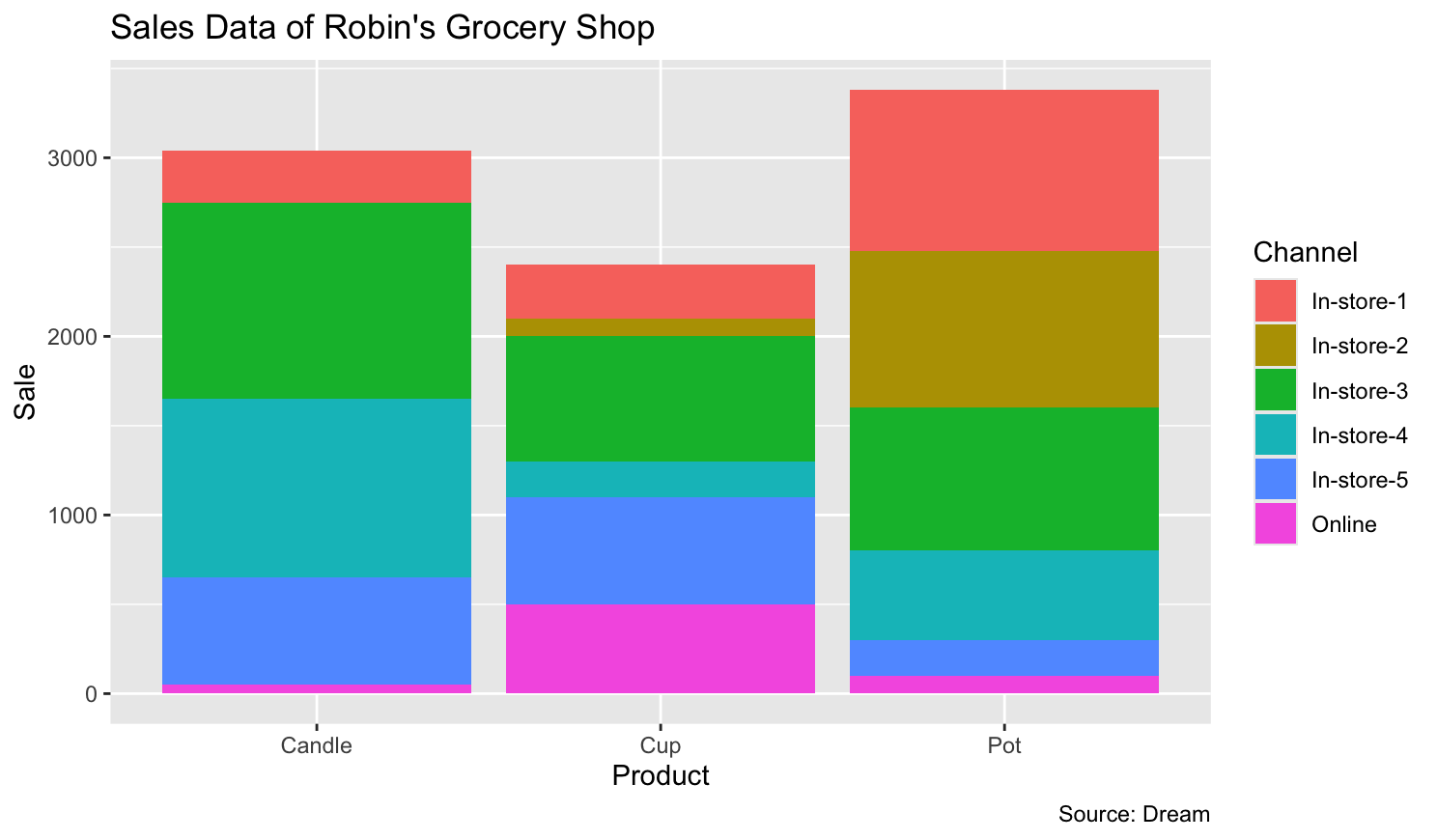


Figure 5 Robin's Grocery sales data

Therefore, when presenting the yearly proportion of the major news categories, only three of the most frequent news were chosen each year to prevent the issue of visual clutter, and the categories left were concluded as a new single group named ‘other’. Although bar chart is often used to visualize categorical data by showing the frequency of each group, the proportion of different types of news were plotted in the present stacked bar chart due to the aim of this visualization is to explore the most popular news with a related higher proportion to other news categories recently (i.e. political news, as shown in the Figure 4.) Also, this scales the y axis to the same value, which is better for comparison at the same level. It also allow people to combine categorical and temporal data visualization, by showing the year in the x axis and the percentage of news types in the y axis, without adding additional dimensions(Von Landesberger et al., 2012).

Alternatively, pie chart is also an effective kind of graph to show the proportion of different categories of a whole and especially useful in comparing the within group difference (Spence & Lewandowsky, 1991; Rangecroft, 2003). However, it might become complicated when combining a temporal variable in it. If we want to use a pie chart to show proportional data and add an annual comparison, multiple pie charts will be drawn to achieve this goal, which will inevitably increase the complexity of the visualization and reduce the efficiency and accuracy of comparisons between the same group at different times (Siirtola, 2019).

An area chart, especially the stacked derived form, could also use to show the proportion trend of different news categories. It could provide a strong visual impact on time series analysis with also showing the proportion changes of different news types and emphasizes the excesses of specific attribute (Moritz & Fisher, 2018). However, to be more specific, the present visualization is mainly focus on showing the related significance of different news groups rather than focusing on the excessive part, which a stacked bar chart might be more suitable.

In conclusion, the use of a stacked bar chart is appropriate in this context because it facilitates a comparative analysis of the relative importance of news categories over time. This method simplifies the visualization’s complexity and improves the efficiency of inter-category comparisons with higher interpretability.

Implications and Improvements

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

Figure 6 Mean sentiment across all news categories

The treemap shows the overall sentiment tendency of each news category by their average sentiment scores. Among them, the size of the region represents the number of news headlines in each category, and the colour gradient (from red to green) reflects the emotional polarity, providing a two-layer perspective. The adoption of a colour gradient is red-green colour, which is influenced by natural associations and traffic lights, as the red-green board is highly intuitive (Zeileis et al., 2019). That is, red is often associated with warnings, danger, or negative emotions, while green is associated with a state of safety or positivity. This visualisation is particularly effective in identifying categories with extreme emotional tendency, such as Crime (-0.95) and Style (0.47). As demonstrated by Malele, Letsoalo & Mafu (2022), treemaps are introduced as an effective way to represent the comparison of categorical data, highlighting their advantages in displaying relative size comparison. Also, treemaps could facilitate rapid identification of trends. For example, news in the Health and Entertainment categories generally exhibit a positive emotional tone, while in contrast, categories such as U.S. News, Politics, and Crime tend to demonstrate a negative sentiment. Past literature provides the evidence of using treemaps to present multi-dimensional data. Both Boumaiza (2015) and Sponder & Khan (2017) discuss the benefits of treemaps in the potential to break data down into subgroupings. While Marwah & Thapar (2020) also highlight the similarities between treemaps and word clouds in showing volume and emotion in sentiment analysis.

However, this method has some limitations in the use of treemaps in this particular study. The inclusion of numerous subcategories might result in a cluttered presentation, which may not only reduce the readability but also could leave the audience overwhelmed and cause difficulties in understanding. As Slingsby, Dykes & Wood (2008) explained, if there are too many data categories presented in a treemap (more than 15-20), due to the limitations of human short-term memory, readers may get confused and have to repeatedly check the annotations. This situation is particularly prominent when presenting spatio-temporal data. Therefore, when it is necessary to display an excessive number of categories, the annotations in the visualisation must be marked clearly. In contrast, interestingly, Long et al. (2017) criticise that it is unnecessary for using treemaps presenting datasets with small samples and a small number of categories by emphasising the scalability of treemaps. In addition, the red-green colour palette can create identification barriers for colourblind users, which in turn limits the accessibility of this visualisation. As identified by Wong (2011) and Fairchild (2013), about 8% of the total population has visual constraints, including dichromacy and trichromacy, in Europe. Thus, choosing effective colours for readers with colour vision deficiency is one of the most important aspects in visualisation (Stauffer et al., 2015). Besides, the low precision of comparison in treemaps is also a frequent criticism, notably when the quantity of categories increases (Slingsby, Dykes & Wood, 2008). For example, it is hard to precisely compare subcategories among small groups, such as the quantity of College news and Education news.

To address the limitations and improve the treemap, firstly, news categories with low quantity with similar content could be combined as new groups, such as college and education, green and environment, and science and technology. For the colour accessibility, although replacing the red-green colour gradient with a colourblind-friendly palette, such as palettes from viridis libraries, could be beneficial, this might break the intuitiveness of the visualization. Thus, the corresponding average sentiment score was added under the name of each news category in the treemap to help identify positive or negative emotions.

Conclusion

To conclude, this study explores how news sentiment changes across different categories and time through four visualisations. These charts are designed to reveal whether media is using political news to manipulate readers’ sentiment and thus provoke viewers to think about the objectivity of journalism. Firstly, the stacked bar chart shows the proportional distribution of different news categories over the years, providing context for the overall trend. This chart provides an in-depth discussion on choosing the right visualization method, emphasizing the clarity of the histogram in showing the changes in categories and the proportion of the charts. Secondly, the combination of area and colour in Treemap visually shows the number of news categories and the overview of sentiment tendencies, while Boxplot shows the variability of the sentiment distribution of news categories in more detail. Finally, the Scatter Point-Line Chart focuses on political news, analysing the potential relationship between sentiment trends and important events such as the UK and US elections. Overall, these four charts provide a comprehensive perspective from data distribution, trend analysis to sentiment changes through multi-dimensional and multi-level presentations. By combining theoretical frameworks, accessibility, visual choices, and suggestions for improvement, Grammar of Graphics, visual accessibility design, etc., are applied to improve the logic, persuasiveness, and friendliness of the charts to different audiences.

Reference

Boumaiza, A. (2015). A survey on sentiment analysis and visualization. Journal of Emerging Technologies in Web Intelligence, 7(1).

Brehmer, M., & Munzner, T. (2013). A Multi-Level Typology of Abstract Visualization Tasks. IEEE Transactions on Visualization and Computer Graphics, 19(12), 2376–2385. https://doi.org/10.1109/TVCG.2013.124

Cambridge University Press. (n.d.). News. In Cambridge Dictionary. Retrieved January 20, 2025, from <https://dictionary.cambridge.org/dictionary/english/news>

Evans, M. D., & Lyons, R. K. (2008). How is macro news transmitted to exchange rates? Journal of Financial Economics, 88(1), 26–50.

Fairchild, M. D. (2013). Human color vision. In Color Appearance Models (pp. 1–37). John Wiley & Sons.

Ferster, B. (2013). Interactive visualization: Insight through inquiry. MIT Press.

Few, S. (2009). Now you see it: Simple visualization techniques for quantitative analysis. Analytics Press.

Healey, C. G. (1996). Choosing effective colours for data visualization. Proceedings of Seventh Annual IEEE Visualization ’96, 263–270. https://doi.org/10.1109/VISUAL.1996.568118

Hopmann, D. N., Vliegenthart, R., De Vreese, C., & Albæk, E. (2010). Effects of election news coverage: How visibility and tone influence party choice. Political Communication, 27(4), 389–405. https://doi.org/10.1080/10584609.2010.516798

Huang, C. J., Liao, J. J., Yang, D. X., Chang, T. Y., & Luo, Y. C. (2010). Realization of a news dissemination agent based on weighted association rules and text mining techniques. Expert Systems with Applications, 37(9), 6409–6413.

Indratmo, Howorko, L., Boedianto, J. M., & Daniel, B. (2018). The efficacy of stacked bar charts in supporting single-attribute and overall-attribute comparisons. Visual Informatics, 2(3), 155–165. https://doi.org/10.1016/j.visinf.2018.09.002

Kim, N. W., Joyner, S. C., Riegelhuth, A., & Kim, Y. (2021, June). Accessible visualization: Design space, opportunities, and challenges. Computer graphics forum, 40(3), 173-188.

Kirk, A. (2012). Data visualization: A successful design process (1st edition.). Packt Publishing, Limited.

Kirk, A. (2012). Data visualization: A successful design process (1st edition.). Packt Pub.

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. Information Sciences, 278, 826–840. https://doi.org/10.1016/j.ins.2014.03.096  
Misra, R. (2022). News Category Dataset. Kaggle. https://www.kaggle.com/datasets/rmisra/news-category-dataset?resource=download

Long, L. K., Hui, L. C., Fook, G. Y., & Zainon, W. M. N. W. (2017). A study on the effectiveness of tree-maps as tree visualization techniques. Procedia Computer Science, 124, 108-115.

MacEwan, E. J. (1900). Freytag’s technique of the drama: An exposition of dramatic composition and art. Scott Foresman.

Malele, V., Letsoalo, M. E., & Mafu, M. (2022, August). Sentiment Analysis of South African News Company. In 2022 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD) (pp. 1-5). IEEE.

Marwah, M. S., & Thapar, R. Sentiment Analysis of the New Education Policy 2020: Enhancing Adult Education and Lifelong Learning for Employability.

Moritz, D., & Fisher, D. (2018). Visualizing a million time series with the density line chart. arXiv preprint arXiv:1808.06019.

Munzner, T. (2016). Keynote speaker: Visualization analysis and design. 2016 IEEE Pacific Visualization Symposium (PacificVis), xiii–xiii. https://doi.org/10.1109/PACIFICVIS.2016.7465242

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015). Text mining of news-headlines for forex market prediction: a multi-layer dimension reduction algorithm with semantics and sentiment. Expert Systems with Applications, 42(1), 306–324.

Norman, D. (1988). The design of everyday things. Doubleday.

Norris, P., Curtice, J., Sanders, D., Scammell, M., Semetko, H. A., & Baker, D. (2000). On message: Communicating the campaign [Review of On message: communicating the campaign]. Democratization, 7(2), 231–233.

Peramunetilleke, D., & Wong, R. K. (2002). Currency exchange rate forecasting from news headlines. Australian Computer Science Communications, 24(2), 131–139.

Rangecroft, M. (2003). As easy as pie. Behaviour and Information Technology, 22(6), 421-426.

Sampanes, A. C., Tseng, P., & Bridgeman, B. (2008). The role of gist in scene recognition. Vision research, 48(21), 2275-2283.

Seifollahi, S., & Shajari, M. (2019). Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to FOREX market prediction. Journal of Intelligent Information Systems, 52(1), 57–83. https://doi.org/10.1007/s10844-018-0504-9

Sherry, M. B., Agosto, V., Blank, J., Cain, L., Feldman, A., Jung, K., & Wolgemuth, J. (2021). On Being “a Methodologist” in Five Ws+ H. Research in the Schools, 27(1), 34–43.

Shneiderman, B., & Plaisant, C. (2010). Designing the user interface: Strategies for effective human-computer interaction (5th ed., International ed.). Addison-Wesley.

Siirtola, H. (2019). The Cost of Pie Charts. 2019 23rd International Conference Information Visualisation (IV), 2019, 151–156. https://doi.org/10.1109/IV.2019.00034

Slingsby, A., Dykes, J., & Wood, J. (2008). Using treemaps for variable selection in spatio-temporal visualisation. Information Visualization, 7(3-4), 210-224.

Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. Applied Cognitive Psychology, 5(1), 61-77.

Sponder, M., & Khan, G. F. (2017). Advanced Text Analytics and Algorithms. In Digital Analytics for Marketing (pp. 225-259). Routledge.

Stauffer, R., Mayr, G. J., Dabernig, M., & Zeileis, A. (2015). Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations. Bulletin of the American Meteorological Society, 96(2), 203-216.

Streit, M., & Gehlenborg, N. (2014). Points of View: Bar charts and box plots. Nature Methods, 11(2), 117. https://doi.org/10.1038/nmeth.2807

von Landesberger, T., Bremm, S., Andrienko, N., Andrienko, G., & Tekusova, M. (2012). Visual analytics methods for categoric spatio-temporal data. 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), 183–192. https://doi.org/10.1109/VAST.2012.6400553

Wong, B. (2011). Color blindness. Nat. Methods, 8, 441. doi:10.1038/nmeth.1618.

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., … & Wilke, C. O. (2019). colorspace: A toolbox for manipulating and assessing colors and palettes. arXiv preprint arXiv:1903.06490.