The Journal of Open Source Software

SIRITVIS: Social Interaction Research Insights Topic Visualisation

Sagar Narwade [1], **Gillian Kant** [2], **Benjamin Säfken** [1], **and Benjamin Leiding**[1]

1 Technische Universität Clausthal, Clausthal-Zellerfeld, Germany 2 Georg-August-Universität Göttingen, Göttingen, Germany

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

SIRITVIS is a comprehensive text analysis tool developed to analyze data from Reddit, Instagram, and various external text data sources. It utilizes advanced methodologies to extract information, clean data, and optimize topic modeling efficiently. The package includes data visualization tools that assist in exploring and understanding textual data. A notable feature of SIRITVIS is its ability to map social media posts globally, correlating geographical locations with trending topics, thus providing insights into global trends and digital conversations. It also performs sentiment analysis on mapped data using the VADER tool (Hutto & Gilbert, 2014). This software is a valuable resource for the scientific community, enabling the exploration of public discussions across multiple social platforms. It supports a range of research purposes, including the analysis of current discussions on global geopolitical issues. The installation of SIRITVIS is straightforward and facilitated by the pip package manager. Detailed installation instructions are available in the package's repository[1].

Statement of Need

The proliferation of social media platforms has significantly changed how we communicate, share information, and express opinions on a variety of topics. Platforms like Reddit and Instagram serve as major hubs for public discussion. Analyzing text data from these platforms can provide insights into public sentiments, preferences, and trending discussions, which are valuable for areas such as marketing, politics, and disaster management.

Handling the large volume of unstructured text data from social media can be challenging due to its dynamic and expansive nature. To help with this, we present SIRITVIS, an open-source text analysis package designed to facilitate the exploration and analysis of social media data. SIRITVIS incorporates several advanced neural topic models, including AVITM (Srivastava & Sutton, 2017), as well as other widely used models like Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Neural Latent Dirichlet Allocation (NeuralLDA) (Srivastava & Sutton, 2017), Product Latent Dirichlet Allocation (ProdLDA) (Srivastava & Sutton, 2017), and Contextualized Topic Models (CTM) (Bianchi et al., 2021). These models help in automatically identifying and extracting key topics in an unsupervised manner, allowing users to efficiently explore large text datasets and uncover meaningful patterns.

SIRITVIS aims to provide users with a tool that simplifies the complex task of analyzing social media text data, making it accessible and practical for a variety of applications.

SIRITVIS is an open-source toolset designed for extracting and analyzing data from social media platforms, including Reddit and Instagram, using their respective APIs. The software facilitates a seamless process of data extraction, followed by detailed preprocessing, where

[1] https://github.com/CodeEagle22/SIRITVIS/

essential information is distilled from raw data and extraneous elements are removed through advanced natural language processing (NLP) techniques. This processed data is then used for topic modeling, with users having the option to adjust hyperparameters based on their specific needs and domain expertise.

The package also features an evaluation module that allows users to assess trained models using a variety of metrics that can be customized to fit particular analytical requirements. Additionally, SIRITVIS includes functionalities for analyzing and retrieving the most proficiently trained models, enhancing its utility for research and analysis.

To further aid in the understanding and interpretation of textual data, SIRITVIS incorporates two robust data visualization tools: PyLDAvis and Word Cloud. PyLDAvis (Sievert & Shirley, 2014), enables a more accessible interpretation of topic models derived from textual data by creating dynamic and interactive visualizations that help users explore the relationships between topics and their associated keywords (see Figure 1). The Word Cloud tool (Mueller, 2023), provides a simple yet effective way to visually represent the most frequently occurring words in a dataset, making it easier to identify key terms and patterns at a glance (see Figure 2). These tools collectively offer a comprehensive and intuitive approach for uncovering and communicating the hidden patterns and insights within textual data.

SIRITVIS is renowned for its remarkable ability to map the spatial distribution of Instagram posts and Reddit comments onto a global map, linking each geographical location with its top trending topics and their respective frequencies (see Figure 3). Moreover, it excels in the color-coding of these locations based on the sentiments expressed in each post, providing an accurate count of positive, negative, and neutral posts (see Figure 4). Furthermore, SIRITVIS facilitates a user-friendly exploration of specific keywords and visualizes their occurrences on the world map. This spatial insight contributes significantly to an enhanced understanding of public discussions and lends invaluable support to data-driven decision-making across diverse domains.

Comparing and Contrasting Available Toolsets

In recent years, the field of text data analysis from social media has seen significant advancements, offering researchers a variety of tools and approaches to analyze the wealth of online content. Within this evolving landscape, the SIRITVIS framework offers a unique approach that differentiates itself from existing tools such as TTLocVis (Kant et al., 2020), TweetViz (Stojanovski et al., 2014), and Twitmo (Abuchmueller, n.d.).

SIRITVIS provides a flexible, user-friendly, and comprehensive solution for analyzing social media text data. It includes a wide range of advanced topic models and a unique capability for identifying geographical information. Additionally, SIRITVIS integrates seamlessly with pyLDAvis (Sievert & Shirley, 2014), enabling users to visualize the outcomes of their analyses effectively.

A distinguishing feature of SIRITVIS is its comprehensive suite of evaluation metrics, supported by the octis tool (Terragni et al., 2021). These metrics include topic diversity, accuracy, inverted RBO, coherence, and Jaccard similarity, ensuring that the topic models generated are both reliable and meaningful. This robust evaluation framework reflects a commitment to producing substantive and high-quality results.

SIRITVIS also emphasizes ease of use and accessibility. Its intuitive interface and detailed documentation are designed to accommodate both novice and experienced users, facilitating ease of navigation and application. Moreover, SIRITVIS is compatible with various data sources and formats, allowing researchers to tailor it to their specific needs without difficulty. This focus on accessibility broadens its utility, making advanced text data analysis more accessible to a wider range of researchers and practitioners.

The Journal of Open Source Software

Figures

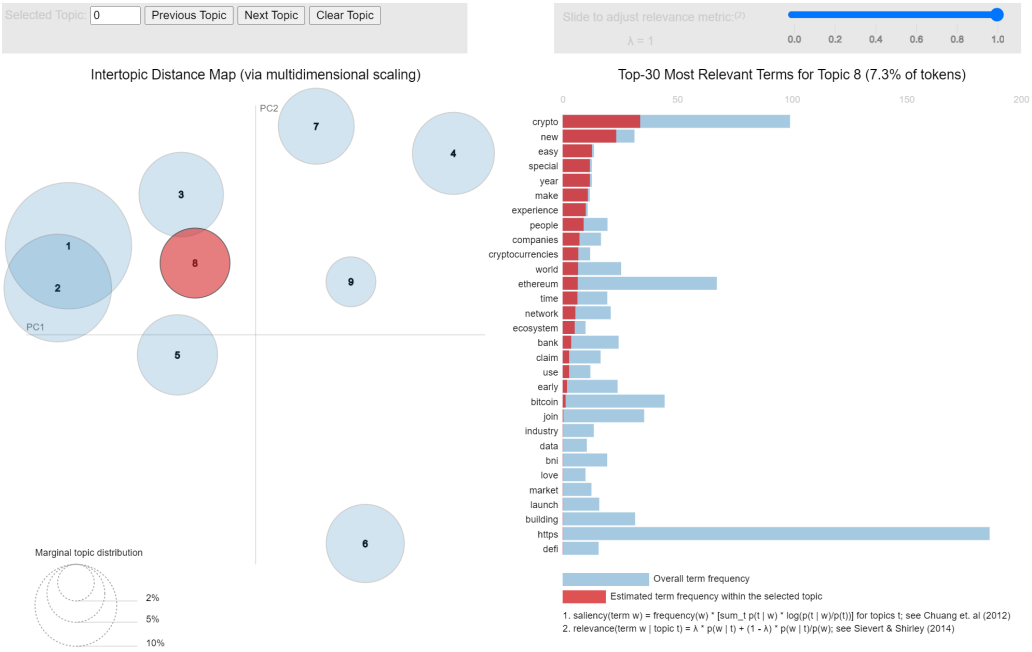

Figure 1: pyLDAvis Visualisation

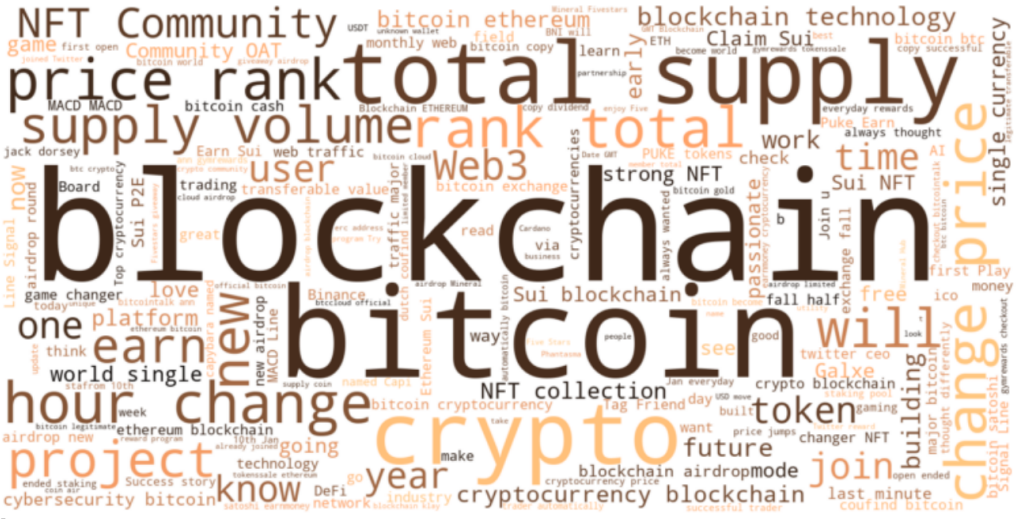

Figure 2: Word Cloud

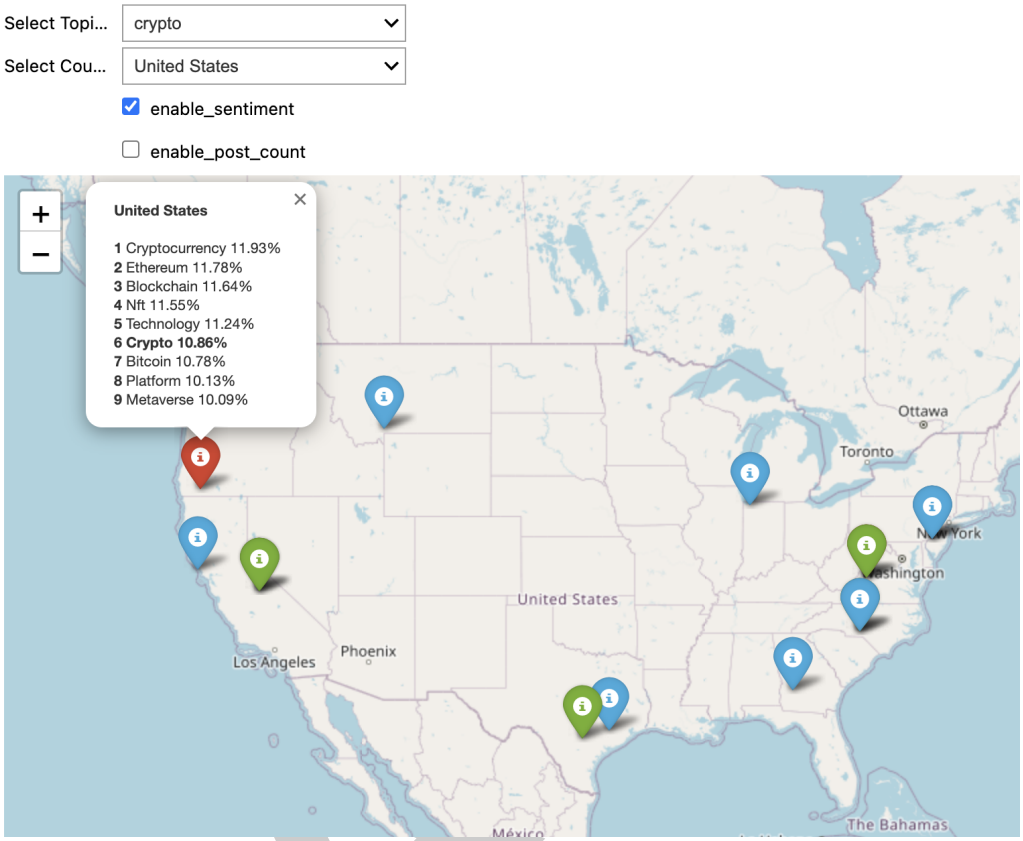

Figure 3: Topic Mapper Frequency Count

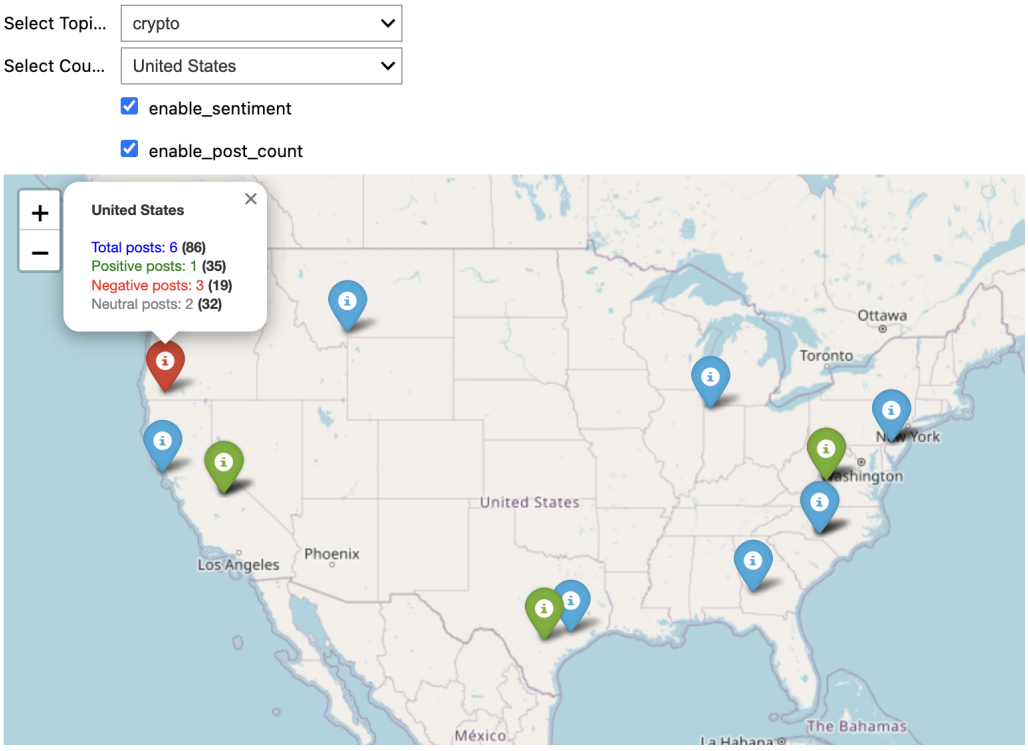

Figure 4: Topic Mapper Sentiment Count

The Journal of Open Source Software

Acknowledgements

We wish to formally acknowledge and express our gratitude for the substantial contributions of Christoph Weisser and Michael Schlee throughout the entire duration of this project.

References

Abuchmueller. (n.d.). *GitHub - abuchmueller/Twitmo: Collect Twitter data and create topic models with R*. https://github.com/abuchmueller/Twitmo

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2021). Cross-lingual contextualized topic models with zero-shot learning. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1676–1683. https://doi.org/10.18653/v1/2021.eacl-main.143

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Kant, G., Weisser, C., & Säfken, B. (2020). TTLocVis: A twitter topic location visualization package. *Journal of Open Source Software*, *5*(54), 2507. https://doi.org/10.21105/joss.02507

Mueller, A. C. (2023). *Wordcloud* (Version 1.9.1). https://github.com/amueller/wordcloud

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In J. Chuang, S. Green, M. Hearst, J. Heer, & P. Koehn (Eds.), *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3110

Srivastava, A., & Sutton, C. (2017). *Autoencoding variational inference for topic models*. https://arxiv.org/abs/1703.01488

Stojanovski, D., Dimitrovski, I., & Madjarov, G. (2014). Tweetviz: Twitter data visualization. *Proceedings of the Data Mining and Data Warehouses*, 1–4.

Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., & Candelieri, A. (2021). OCTIS: Comparing and optimizing topic models is simple! In D. Gkatzia & D. Seddah (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: System demonstrations* (pp. 263–270). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-demos.31