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Published: unpublished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 powerful text analysis tool that has been carefully designed to analyse data from Reddit, X (Twitter), Instagram, and any external text data sources. It utilises the latest tools to automatically extract information from these sources, clean the data, and optimise topic modeling more efficiently. The Package also contains data visualisation tools to improve the investigation and understanding of textual data. One feature of SIRITVIS is its ability to map social media posts on a global scale, connecting geographical locations with trending topics, thus providing valuable insights into the worldwide trends and conversations shaping our digital landscape. Additionally, it offers sentiment analysis on mapped data using the well-regarded VADER tool (Hutto & Gilbert, 2014). This software is a valuable resource for the scientific community, offering deep insights into public discussions on various social platforms. It can be used for a wide range of research purposes, including the analysis of recent discussions on global geopolitical issues. Installing SIRITVIS is straightforward, as it can be accomplished using the pip package manager. Comprehensive installation instructions can typically be found in the dedicated repository of the package.[1]

Statement of Need

The surge of social media platforms has revolutionised the way we communicate, share information, and express viewpoints on various subjects. Among these platforms, Reddit, Instagram, and X (Twitter) stand out as notable sources of public discourse. Examining text data from these platforms offers valuable insights into public sentiments, preferences, and trending discussions, benefiting fields such as marketing, politics, and disaster management.

Dealing with the colossal volume of unstructured text data from social media is challenging due to its dynamic nature and sheer size. To address this challenge, we introduce SIRITVIS, a text analysis package designed to simplify the analysis of social media data. The package employs advanced neural topic models developed by AVITM (Srivastava & Sutton, 2017) and other popular topic models, including 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 Contextualised Topic Models (CTM) (Bianchi et al., 2021). These models automatically identify and extract top topics in an unsupervised manner, enabling users to explore large text datasets and discover hidden patterns for meaningful insights.

SIRITVIS encompasses a comprehensive suite of functionalities designed for data extraction from social media platforms, including X (Twitter), Reddit, and Instagram, facilitated through the utilisation of respective application programming interfaces (APIs). This extraction process

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

is followed by a meticulous data preprocessing phase, wherein valuable information is extracted from raw data, and superfluous elements are expunged, utilising advanced natural language processing (NLP) techniques. The resultant processed data is subsequently employed for topic modeling, with the flexibility for users to fine-tune hyperparameters according to their domain expertise.

In addition, the package includes an evaluation module, which allows for the assessment of trained models using a variety of metrics tailored to the user's specific requirements. Furthermore, the software offers the capability to analyse and retrieve the most proficiently trained models, further enhancing its utility and practicality for scientific research and analysis.

To enrich the analysis and comprehension of textual data, SIRITVIS includes two powerful data visualisation tools: PyLDAvis (Sievert & Shirley, 2014) and Word Cloud (Mueller, 2023). The useful data visualisation tool PyLDAvis, created by Sievert and Shirley, improves the interpretation of topic models from textual data. In order to give users a deeper and more intuitive understanding of the latent themes contained in the text corpus, it creates dynamic and interactive visualisations that assist users in exploring the connections between subjects and their associated keywords (see figure 1). On the other hand, the Word Cloud tool offers an engaging and straightforward means of visually representing the most frequently occurring words in the dataset, simplifying the identification of crucial keywords and patterns at a glance (see figure 2). Together, these tools provide users with a comprehensive and user-friendly approach to uncover and communicate the hidden patterns and insights within their textual data.

SIRITVIS is renowned for its remarkable ability to map the spatial distribution of tweets and Instagram posts 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 visualises 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 witnessed remarkable advancements, offering researchers and practitioners an array of toolkits and approaches to delve into the wealth of online content. Within this dynamic landscape, it becomes imperative to discern the distinctive features of our research, encapsulated in the SIRITVIS framework, as it stands apart from existing related work.

Although alternatives such as TTLocVis (Kant et al., 2020), TweetViz (Stojanovski et al., 2014) and Twitmo (Abuchmueller, n.d.) have their merits, SIRITVIS sets itself apart by providing exceptional flexibility, usability, and comprehensiveness. Its extensive array of advanced topic models, alongside a distinctive capability for pinpointing geographical information and seamless integration with pyLDAvis (Sievert & Shirley, 2014) for visualising outcomes, empowers researchers to extract profound insights from social media text data.

What sets SIRITVIS apart is its comprehensive suite of evaluation metrics, facilitated by the octis tool (Terragni et al., 2021). These metrics cover important aspects such as topic diversity, accuracy, inverted RBO, coherence, and Jaccard similarity, ensuring that the topic models generated by SIRITVIS are not only reliable, but also imbued with substantive meaning. This robust evaluation framework is a hallmark of the research, emphasizing the toolkit's dedication to producing meaningful results.

Furthermore, SIRITVIS places a strong emphasis on user-friendliness and accessibility. Its intuitive interface and detailed documentation cater to both novice and experienced users, making the toolkit approachable and easy to navigate. Additionally, SIRITVIS is designed to

accommodate various data sources and formats, ensuring that researchers can adapt it to their specific research needs effortlessly. This commitment to accessibility broadens the toolkit's reach, making advanced text data analysis more accessible to a wider community of researchers and practitioners.

Figures

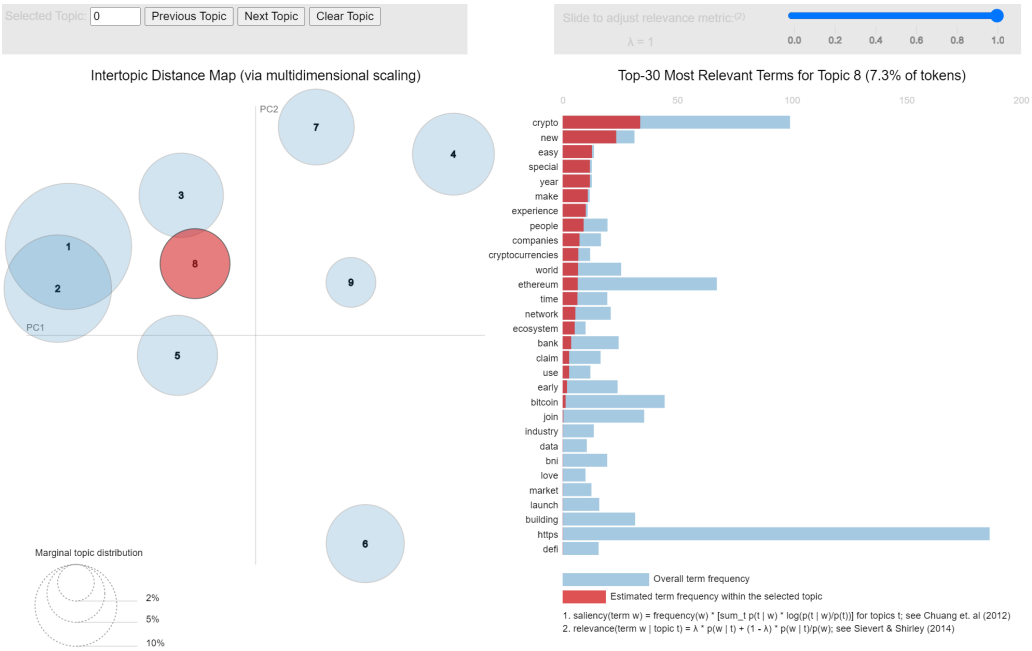

Figure 1: pyLDAvis Visualisation

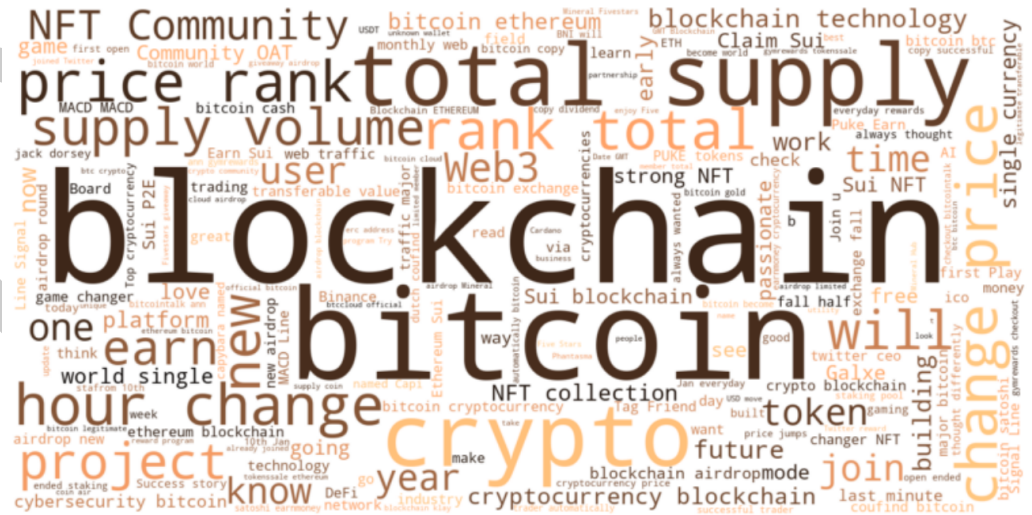

Figure 2: Word Cloud

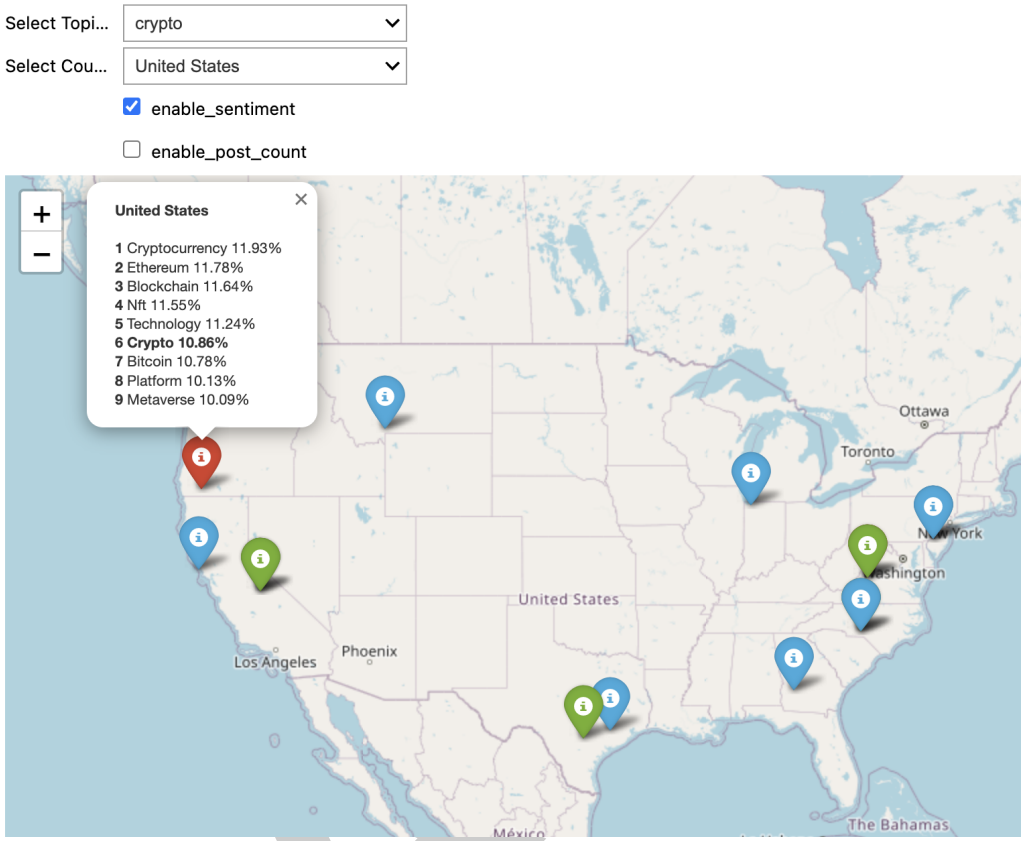

Figure 3: Topic Mapper Frequency Count

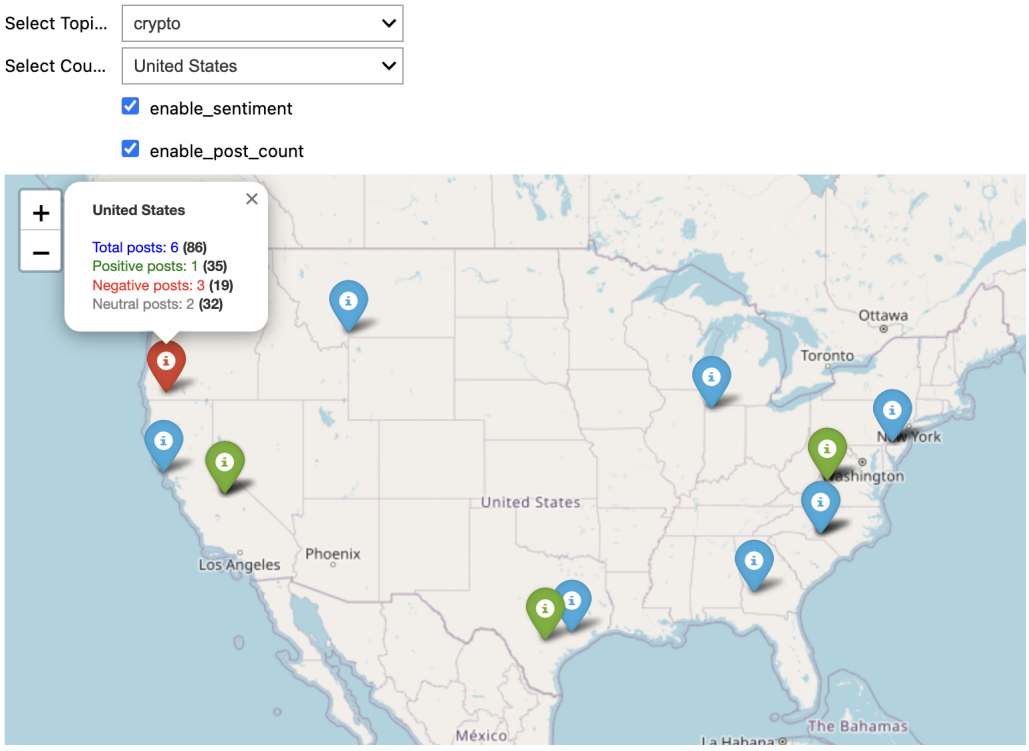

Figure 4: Topic Mapper Sentiment Count

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