Prototyping Text Mining and Network Analysis Tools to Support Netnographic Student Projects

https://doi.org/10.3991/ijet.v15i10.12313

Ilya Musabirov (✉), Denis Bulygin
National Research University Higher School of Economics, Saint-Petersburg, Russia
ilya@musabirov.info

Abstract—Social science is witnessing tremendous growth of data available on the Internet regarding social phenomena; however, social science students are typically not prepared for managing the challenges and opportunities of analysing online data. One of the areas where this growth is especially important is in social studies of consumption. This article discusses a prototype of a visualisation tool intended to support the learning of netnographic analysis with computational tools.

Keywords—Netnography, valuation studies, virtual consumption, education tools, visualisation

1 Introduction

One of the important aspects of applying new skills for social science students is working with online data that provide valuable opportunities to learn more about social phenomena [1]. In fields such as the sociology of consumption, valuation studies, and marketing research, the text data of online discussions can be a fruitful source of information about the motivations of people, their perceived dimensions of experience, and evaluation practices. However, qualitative or manual quantitative content analysis of large corpora becomes very demanding in terms of time and other resources [2].

This paper presents a design concept of a visualisation tool that allows students to analyse user-generated content from a netnographic perspective. By combining activity-centred design (ACD) [3], [4] and instructional design practices, we prototype this software tool that covers the common research goals of netnographers.

By creating a visualisation tool, we aim to support the exploration of data collections using text mining and network analysis methods. Moreover, the tool is aimed to help students structure the exploration process and connect findings to theoretical concepts using affordances provided by the structured visualisation of data, facilitating a balance between building quantitative summaries of discussions and conducting a deeper qualitative understanding of texts.

In the article, we demonstrate the prototype on the case of esports athletes brand analysis [5].

2 Netnographic Approach in Studies of Consumption

The development of Internet technologies has led to the emergence of new sources of user-generated data that allow researchers to study different aspects of consumption. Internet users now have various platforms, such as blogs [6], social media [1], [6], [7], review-based websites [8]–[10], and websites of organisations and public institutions [11], to express themselves and share opinions about their experiences. In this way, user-generated content (UGC) provides researchers with data on what users consume and how they evaluate the consumption process.

Netnography [7] is one of the contemporary approaches to studying online communities, employing a mixed method that is targeted to build an understanding of users and their interactions. Netnography can help online studies answer questions associated with traditional ethnographic face-to-face studies [12] that are usually applied to real-life situations when actions and only visible interactions are examined and are not fit for working with massive concentrations of text data. Netnography helps to focus on cases where text and other forms of communication prevalent in online platforms are the main mediums of meaning. Netnographic quantitative analysis reveals interesting patterns and highlights the most important pieces of materials. By using quantitative methods, netnography supports qualitative analysis and navigates the researcher through the large corpus of data. This analysis can also provide insights into consumer behaviour in particular cases, equipping social science students with tools to bridge the gap between theoretical models they study and applied analytics [13].

3 Computational Methods to Support Netnographic Analysis

Our prototype is based on tools and methods adopted from the emerging areas of computational social science and digital humanities to support netnographic analysis of large UGC datasets and provide students with a repertoire of methods and practices combined in a single interface and blended in learning design. At this stage of the project, we build on the approaches used to analyse text data on different levels, starting from simple word-frequency-based methods to topic modelling and network analysis of concepts and entities extracted from UGC.

As texts are usually the primary content of netnographic studies, computational support of text analysis is the key component of our prototype. Methods based on word-frequency analysis and derivative weights and metrics (e.g., TF-IDF, log-likelihood ratio-based methods) allow us to search for the texts that are most relevant to particular queries, quickly grasp the meaning of texts, and analyse the changes in the corpora. In combination with specialised dictionaries, reflecting sentiment or particular topics, more nuanced analysis is possible.

However, analysing the context of discussions, which is one of the main requirements for netnographic analysis, requires more advanced methods, ranging from n-grams (commonly used token collocations of length n) to topic models and advanced deep learning models.

In our example case, we derived bigrams from text and applied log-likelihood ratio-based comparison to the spectators' discussions of players before and after team changes.

Topic modelling [14], [15] allows researchers to find themes that show consumers' reflections on their experiences. It is a machine-learning-based method of fuzzy bi-clustering of words (n-grams) and documents into groups called topics. When applying this method, researchers ignore all relationships between words in the text: e.g., neighbourship, the position of the words in the text, and lexical meaning. In this method, the text becomes a bag-of-words that only provides the model with information about the co-occurrence of words. Considering the frequency of words and their co-occurrence, the model creates topics and sequences of unique words distributed by their probabilities related to the same group. The role of the researcher in topic modelling is twofold. First, the researchers define the number of topics that should be created during topic model computation. By relying on diagnostic metrics [16], [17], the researcher can make such a decision. Second, the researcher should interpret the groups of words in an attempt to reveal the themes. By using the list of most probable words and examples of text with the highest proportion of a topic, the researcher can label each topic. While other advanced methods of text mining exist, showing superior performance on many natural language processing tasks, topic modelling balances performance and interpretability, making it a suitable and widespread instrument for social science goals [18].

Another approach used in the prototype is network analysis [18] of relationships between the entities mentioned in the text. Network analysis is a fruitful method of analysing relationships between tokens in the texts. What can the co-occurrence of tokens in the text tell us about community practices? In discussions of brands, community members compare athletes and mention their names in the posts, allowing us to visualise the network of brands' co-mentioning. In addition, networks are useful when scholars are focused on understanding similarities between compared brands and the contexts of discussion that mention two brands – i.e., what users write in the forums when mentioning two brands. For example, is co-mentioning a case of finding dis-similar brands or stating that one brand is better than another?

4 Design Methods

There are already solutions to help scholars to code and visualise ethnographic data (e.g., [19]); however, those tools are less suitable for supporting the analysis of large bodies of text data, which is an important task of netnographic studies. Our approach implies rapid prototype design and development, building on existing open-source software components in the R ecosystem [20] and modern visualisation and interactive presentation tools.

Within the spirit of activity-centred design, we use the Jobs-to-be-done framework [21] to focus on representing users' needs in a wider netnographic analysis context. Afterwards, we map netnographers' activities with learning goals to give students competences to perform these activities with computational tool support.

Because the focus is on building the tool supporting learning real research activities, we rely on building the prototype based on the available algorithms, methods, and software instruments belonging to the R project ecosystem [20].

This approach allows for the gradual removal of tool-provided scaffolding for those students progressing to a deeper technical level of expertise and fast integration of the new methods and activities.

There are existing approaches to support exploration for some of the tasks. For example, there is a tool for topic modelling results exploration called LDAV, which introduces a novel metric for presenting words connected with topics and model tuning [22]. However, these are task-focused and not activity-focused, thus making them less relevant to support both research and learning goals. Instead, we use computational instruments to perform computational text analysis [23] and network analysis [24], and we rely heavily on visualisations and scaffolds in the form of applied recipes, parameters, and connections between different instruments and stages of analysis.

For visualisations, we use ggplot2 ecosystem packages [25], [26] built on using a grammar of graphics approach. Regarding the framework for the application, we use a flexdashboard [27] package and an interactive web-service package called Shiny [28].

We believe that our approach helps to embrace a more holistic perspective in our design, interconnecting preprocessing, interpretation, and results construction parts of computationally supported netnographic analysis.

5 Activities and Learning Tasks

Based on the literature review and syllabi analysis of related courses and projects, we attempted to extract the main activities that netnographers perform during work with collected text data (Table 1). Then, we connected these activities with learning goals extracted from the course and project syllabi and mapped the supporting computational technologies and related concepts.

Table 1. Examples of job stories and learning goals

#	Job Story	Learning goals
1	I want to Know themes of discussions and assess their prevalence So that I can Understand what users are talking about and what they pay more attention to	Interpret discussions using the results of the computational model Choose the texts representing the important patterns in discussions Visualise the structure of text collection and its connection to codes
2	I want to Know the relationships between several entities mentioned in discussions So that I can Learn how community participants connect entities in discussions	Extract entities from the text Decide on the strength of relationships Visualise relationships between entities
3	I want to Compare discussions at several points of time or in sub-communities So that I can Understand the dynamics of discussion or differences between sub-communities	Reveal differences between sub-corpora on a statistical level Visualise revealed differences

By using the four-component instructional design approach [29], [30], we construct a sequence of learning tasks built around authentic netnographic activities supported by part-task practice and procedural and supporting information. In our case, the task sequence starts from a focus on instrument-related skills with a high level of scaffolding. Then, the sequence progresses on core tasks focusing on the interpretation of project data in a netnographic sense, with supporting information communicating details of netnographic approaches and procedural information communicating necessary details about tool applications. Lastly, the focus of the tasks switches to higher-level theoretical, methodological, and ethical problems. While core methodological assumptions of the approach are introduced in [7] and further discussed in multiple studies [31], the challenging tasks for advanced students can involve reflections on the traits of big data and are based [32] and structured around concepts such as algorithmic confoundedness, system, usage, and population drifts.

6 Design and Prototype

We build a prototype with dashboard-based representations, mapping each representation to the supported activity. In this section, we present the prototypes for two activities: analysing themes in the text (see Fig. 1) and analysing relationships between entities in text.

6.1 Themes in text

When working with a large body of text, the vital task for the netnographer is to correctly interpret what participants discuss on social media. This task includes an understanding of the main discussion topics and an evaluation of topics' prevalence and comparison of discussions (e.g., a discussion at several points of time or discussions in several communities). As mentioned previously, topic models produce two kinds of outputs, and the distribution of topics by their proportion in each text and the distribution of words by their probability are related to each topic. By being able to select the most probable words and the texts with the prevalent topic of interest, netnographers can interpret topics.

In interpreting topics, we can rely on different simple probability-based metrics or more complex ones, such as FREX (FRequency + EXclusivity) [23]. The FREX score combines the frequency of a word and its exclusivity in the topic that defines a measure of a word to be related exclusively to a particular topic. The interpreter should rely on multiple scores and their comparisons in the process of interpretation.

Once the topic model is calculated and topics are labelled, researchers can use topic proportions to apply the methods of statistical inference and qualitative analysis of web community content. In the spirit of netnography [7], [33], it is possible to use the topic model as a snapshot of discussions that reveals key themes and highlights the most interesting texts. Based on this approach, researchers treat topics as frames of discussions [2], and texts with the highest proportion of a particular topic can demonstrate in what context consumers mention particular themes and how texts on

this theme can vary. Moreover, netnographers that deal with several corpora of texts can quickly compare their content, examining the tokens' prevalence in one of the corpora.

6.2 Relationships between entities of interest in texts

Visualisation of networks allows researchers to learn more about how entities under study are connected. Relying on information about connections between entities in texts helps researchers understand who or what are mentioned together in discussions. By building on this baseline structure, netnographers can start exploring attribute-based properties of the network and mapping attributes to visualisations. Network attribute distribution for whole or personal networks for some entities can be summarised with a radar chart, allowing us to make comparisons and build profiles.

While in traditional social research, personal networks are most often self-reported, in netnographic studies, they can be reconstructed based on the discussions. In our example case, we are representing community perceived connections between a particular athlete and the other athletes with which they are associated in texts, and analyse which attributes they have in common, suggesting possible explanations for the comparisons and allowing the researcher to reason about possible mechanisms behind them, and support this reasoning with relevant examples.

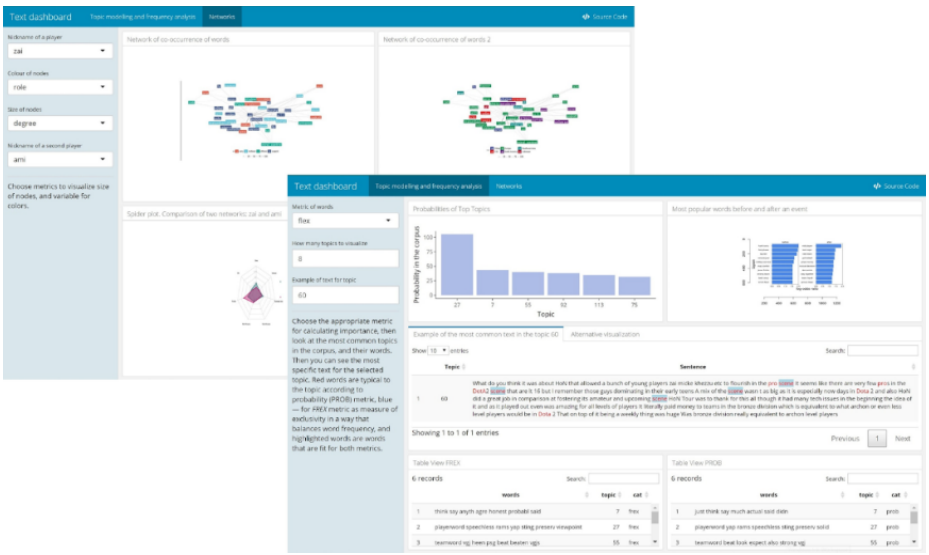

Fig. 1. Screen with contextual information for analysis of topics

7 Conclusion and Future Work

While we describe only the first results of the proposed approach, some current directions for software support are clear, and improving the program's ease of use and

automating typical operations will help lower the entrance barrier for undergraduate students.

When supporting progression to more professional use of the computational instruments, resulting recipes and practices can be detached from the current interface and be reused, e.g., for integration with existing high-level tools such as jamovi [34] or R-QDA [35].

Further challenges are associated with investing in the usability of big text data exploration, focusing on the aspects beyond the scope of traditional-scale content analysis [36] – e.g., ranking cases based on transparent interest measures – and visualising the uncertainty of quantitative estimates [37] and plot layouts to minimise algorithmic confoundedness of explanations produced with the help of computational tools [38]. The current developments of computational data analysis workflow research, the focus on the implementation of visualisation techniques in education [42], [43], and the goals behind these approaches are relevant to enhancing the reproducibility, transparency, and collaboration of netnographic analysis among students. Lastly, recent developments in the interpretability of word-embedding-based [44] and aspect-extraction-orientated NLP methods [45] build a foundation for further integration with computationally supported netnographic research.

8 Acknowledgements

The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2017–2018 (grant No. 18-01-0002) and by the Russian Academic Excellence Project "5-100".

We would like to thank Ekaterina Marchenko for her help.

9 References

[1] J. A. Sandlin, "Netnography as a consumer education research tool," International Journal of Consumer Studies, vol. 31, no. 3, pp. 288–294, May 2007, https://doi.org/10.1111/j.1470-6431.2006.00550.x.

[2] P. DiMaggio, M. Nag, and D. Blei, "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding," Poetics, vol. 41, no. 6, pp. 570–606, 2013. https://doi.org/10.1016/j.poetic.2013.08.004

[3] D. A. Norman, "Logic versus usage: the case for activity-centered design," interactions, vol. 13, no. 6, pp. 45–ff, 2006.

[4] A. Williams, "User-centered design, activity-centered design, and goal-directed design: a review of three methods for designing web applications," in Proceedings of the 27th ACM international conference on Design of communication, 2009, pp. 1–8. https://doi.org/10.1145/1621995.1621997

[5] I. Musabirov, D. Bulygin, and E. Marchenko, "Personal Brands Of Esports Athletes: An Exploration Of Evaluation Mechanisms," National Research University Higher School of Economics, HSE Working papers WP BRP 90/SOC/2019, 2019. https://doi.org/10.2139/ssrn.3501522

[6] R. V. Kozinets, "The Field behind the Screen: Using Netnography for Marketing Research in Online Communities," Journal of Marketing Research, vol. 39, no. 1, pp. 61–72, Feb. 2002, https://doi.org/10.1509/jmkr.39.1.61.18935.
[7] R. V. Kozinets, Netnography: redefined, 2nd edition. Thousand Oaks, CA: Sage Publications Ltd, 2015.
[8] F.-M. Belz and W. Baumbach, "Netnography as a Method of Lead User Identification," Creativity and Innovation Management, vol. 19, no. 3, pp. 304–313, Aug. 2010, https://doi.org/10.1111/j.1467-8691.2010.00571.x.
[9] N. Kaspruk, O. Silyutina, and V. Karepin, "Hotel Value Dimensions and Tourists' Perception of the City. The Case of St. Petersburg," in Digital Transformation and Global Society, 2017, pp. 341–346. https://doi.org/10.1007/978-3-319-69784-0_29
[10] Y. Guo, S. J. Barnes, and Q. Jia, "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation," Tourism Management, vol. 59, pp. 467–483, Apr. 2017, https://doi.org/10.1016/j.tourman.2016.09.009.
[11] S. Linek, A. Schafrick, and K. Tochtermann, "Just for the image? The impact of Web 2.0 for public institutions," International Journal of Emerging Technologies in Learning (iJET), vol. 8, no. 2013, 2013. https://doi.org/10.3991/ijet.v8is1.2266
[12] J. Xun and J. Reynolds, "Applying netnography to market research: The case of the online forum," Journal of Targeting, Measurement and Analysis for Marketing, vol. 18, no. 1, pp. 17–31, 2010. https://doi.org/10.1057/jt.2009.29
[13] V. Ahuja and S. Alavi, "Using Facebook as a Digital Tool for Developing Trust amongst Consumers using Netnography and Social Media Analytics: A Study of Jet Airways," Journal of Relationship Marketing, vol. 17, no. 3, pp. 171–187, 2018. https://doi.org/10.1080/15332667.2018.1440145
[14] D. M. Blei, "Probabilistic topic models," Communications of the ACM, vol. 55, no. 4, pp. 77–84, 2012.
[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.
[16] J. Silge, "Julia Silge - Training, evaluating, and interpreting topic models," 2018. [Online]. Available: https://juliasilge.com/blog/evaluating-stm/. [Accessed: 29-May-2019].
[17] P. Grajzl and C. Irby, "Reflections on study abroad: a computational linguistics approach," Journal of Computational Social Science, pp. 1–31, 2018. https://doi.org/10.2139/ssrn.3235551
[18] P. DiMaggio, "Adapting computational text analysis to social science (and vice versa)," Big Data & Society, vol. 2, no. 2, p. 2053951715602908, 2015. https://doi.org/10.1177/2053951715602908
[19] L. Richards, Using NVivo in qualitative research. Sage, 1999.
[20] "R: The R Project for Statistical Computing." [Online]. Available: https://www.r-project.org/. [Accessed: 10-Nov-2019].
[21] A. Klement, When Coffee and Kale Compete. NYC Publishing, New York, 2016.
[22] 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, 2014, pp. 63–70. https://doi.org/10.3115/v1/w14-3110
[23] M. Roberts, B. Stewart, D. Tingley, and K. Benoit, stm: Estimation of the Structural Topic Model. 2019.
[24] G. Csardi and T. Nepusz, "The igraph software package for complex network research," InterJournal, vol. Complex Systems, p. 1695, 2006.
[25] H. Wickham and W. Chang, "ggplot2: An implementation of the Grammar of Graphics," R package version 0.7, URL: http://CRAN. R-project. org/package= ggplot2, vol. 3, 2008.

[26] T. L. Pedersen, ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. 2018.
[27] R. Iannone et al., flexdashboard: R Markdown Format for Flexible Dashboards. 2018.
[28] "Shiny." [Online]. Available: https://shiny.rstudio.com/. [Accessed: 10-Nov-2019].
[29] J. J. Van Merriënboer and P. A. Kirschner, Ten steps to complex learning: A systematic approach to four-component instructional design. Routledge, 2017. https://doi.org/10.4324/9781315113210-3
[30] J. J. Van Merriënboer, R. E. Clark, and M. B. De Croock, "Blueprints for complex learning: The 4C/ID-model," Educational technology research and development, vol. 50, no. 2, pp. 39–61, 2002. https://doi.org/10.1007/bf02504993
[31] M. Bartl, V. K. Kannan, and H. Stockinger, "A review and analysis of literature on netnography research," International Journal of Technology Marketing, vol. 11, no. 2, p. 165, 2016, https://doi.org/10.1504/ijtmkt.2016.075687.
[32] M. J. Salganik, Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press, 2017. https://doi.org/10.1080/0022250x.2019.1682802
[33] M. Banyai and T. D. Glover, "Evaluating Research Methods on Travel Blogs," Journal of Travel Research, vol. 51, no. 3, pp. 267–277, May 2012, https://doi.org/10.1177/0047287511410323.
[34] "The jamovi project (2019)." [Online]. Available: https://www.jamovi.org/about.html. [Accessed: 10-Nov-2019].
[35] R. Huang, RQDA: Qualitative Data Analysis. 2018.
[36] H. Mei, Y. Ma, Y. Wei, and W. Chen, "The design space of construction tools for information visualization: A survey," Journal of Visual Languages & Computing, vol. 44, pp. 120–132, 2018. https://doi.org/10.1016/j.jvlc.2017.10.001
[37] J. Hullman, "Why Authors Don't Visualize Uncertainty," IEEE transactions on visualization and computer graphics, 2019. https://doi.org/10.1109/tvcg.2019.2934287
[38] C. Xiong, J. Shapiro, J. Hullman, and S. Franconeri, "Illusion of Causality in Visualized Data," IEEE transactions on visualization and computer graphics, 2019. https://doi.org/10.1109/tvcg.2019.2934399
[39] M. B. Kery, B. E. John, P. O'Flaherty, A. Horvath, and B. A. Myers, "Towards Effective Foraging by Data Scientists to Find Past Analysis Choices," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, p. 92. https://doi.org/10.1145/3290605.3300322
[40] A. Head, F. Hohman, T. Barik, S. M. Drucker, and R. DeLine, "Managing Messes in Computational Notebooks," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, p. 270. https://doi.org/10.1145/3290605.3300500
[41] A. Camisetty, C. Chandurkar, M. Sun, and D. Koop, "Enhancing web-based analytics applications through provenance," IEEE transactions on visualization and computer graphics, vol. 25, no. 1, pp. 131–141, 2018. https://doi.org/10.1109/tvcg.2018.2865039
[42] J. Nouri, "Editorial of the First Issue of the International Journal of Learning Analytics and Artificial Intelligence for Education," International Journal of Learning Analytics and Artificial Intelligence for Education (iJAI), vol. 1, no. 1, pp. 4–7, 2019. https://doi.org/10.3991/ijai.v1i1.11073
[43] Á. Hernández-García, I. González-González, A. I. Jiménez-Zarco, and J. Chaparro-Peláez, "Visualizations of online course interactions for social network learning analytics," International Journal of Emerging Technologies in Learning (iJET), vol. 11, no. 07, pp. 6–15, 2016. https://doi.org/10.3991/ijet.v11i07.5889
[44] M. Hurtado Bodell, M. Arvidsson, and M. Magnusson, "Interpretable Word Embeddings via Informative Priors," in Proceedings of the 2019 Conference on Empirical Methods in

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 6323–6329, https://doi.org/10.18653/v1/d19-1661

[45] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "An unsupervised neural attention model for aspect extraction," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 388–397. https://doi.org/10.18653/v1/p17-1036

10 Authors

Ilya Musabirov is a senior lecturer in the Department of Informatics at the National Research University Higher School of Economics St Petersburg. He also works as a junior research fellow at the Sociology of Education and Science Laboratory.

Denis Bulygin is a lecturer in the Department of Informatics at the National Research University Higher School of Economics St Petersburg. He graduated with a BSc degree in Sociology and Social Informatics from the Higher School of Economics, Russia, and an MSc degree in Human–Computer Interaction from Uppsala University, Sweden.

Article submitted 2019-11-11. Resubmitted 2020-02-24. Final acceptance 2020-02-25. Final version published as submitted by the authors.