LaMa: a thematic labelling web application

Victoria Bogachenkova[1*], **Eduardo Costa Martins**[1*], **Jarl Jansen**[1*], **Ana-Maria Olteniceanu**[1], **Bartjan Henkemans**[1], **Chinno Lavin**[1], **Linh Nguyen**[2], **Thea Bradley**[1], **Veerle Fürst**[1], **Hossain Muhammad Muctadir** [1¶], **Mark van den Brand** [1], **Loek Cleophas** [1], and **Alexander Serebrenik** [1]

1 Software Engineering and Technology cluster, Eindhoven University of Technology, Eindhoven, The Netherlands **2** McGill University, Montreal, Canada ¶ Corresponding author * These authors contributed equally.

Summary

Qualitative analysis of data is relevant for a variety of domains including empirical research studies and social sciences. While performing qualitative analysis of large textual data sets such as data from interviews, surveys, mailing lists, and code repositories, condensing pieces of data into a set of terms or keywords simplifies analysis, and helps in obtaining useful insight. This condensation of data can be achieved by associating keywords, a.k.a. *labels*, with text fragments, a.k.a *artifacts*. It is essential during this type of research to achieve greater accuracy, facilitate collaboration, build consensus, and limit bias. LaMa, short for Labelling Machine, is an open-source web application developed for aiding in thematic analysis of qualitative data. The source code and the documentation of the tool are available at https://github.com/muctadir/lama. In addition to being open-source, LaMa facilitates thematic analysis through features such as artifact based collaborative labelling, consensus building through conflict resolution techniques, grouping of labels into themes, and private installation with complete control over research data. With the help of this tool and flow it enforces, thematic analysis becomes less time consuming and more structured.

Statement of need

Analyzing qualitative data has been proven to be labor intensive and time consuming task (Pope et al., 2000) due to its nature. Thematic analysis (Kiger & Varpio, 2020) is a powerful yet flexible method for performing such analysis. Analyzing textual data through this method allows a researcher to understand experiences and thoughts, as well as emotions and behaviors throughout a data set. Due to the flexibility of this analysis method, the users are not bound to using only one paradigmatic perspective but within different data sets can use different ones (Braun & Clarke, 2006).

As thematic analysis is a widely used qualitative analysis technique, several commercial tools are available, such as Atlas.ti[1] and maxQDA[2]. We also investigated open-source applications that allows labelling of artifacts such as Label Studio[3]. Although these tools are very well developed, there are four major trade offs that inspired the development of LaMa.

- **Cost:** The services provided by the commercial tools mentioned above are not free and can be quite expensive depending on the subscription.

---

[1] https://atlasti.com/
[2] https://www.maxqda.com/
[3] https://labelstud.io/

- **Data access and privacy:** Qualitative researches often process sensitive data, such as legally protected information, private information of individuals. With rising privacy concerns, increasing number of research organizations are requiring specialized approval for working with such data. For example, at Eindhoven University of Technology it is mandatory, among other information, to specify which individuals can have access to the research data. With commercial tools, control over the access of the research data or the storage location are often unavailable.

- **Complex collaboration workflow:** Collaborative labelling or coding is an established method for reducing bias during qualitative analysis (Richards & Hemphill, 2017). While commercial tools provide this feature in various forms, the process for resolving conflicting labels is often complicated.

- **Thematic analysis use-case:** We identified several annotation tools, including open-source ones (i.e., Label Studio[4]), that can annotate texts, images, audio and various other data formats. However, these tools are primarily intended to be used together with various machine learning or classification algorithms and therefore, not particularly tailored towards thematic analysis related use-cases.

Based on these points we developed LaMa, which is a web application intended to support the thematic analysis and is built based on an existing application called the Labeling Machine (Muctadir, 2022), which is forked from the earlier Labeling Machine (Aghajani, 2022). In addition to significantly improving the user interface, LaMa provides additional features such as multi-labelling, hierarchical theming, and change tracking. Its key features are described in the following section.

Key features

LaMa in its core functionality is similar to comparable labelling tools for qualitative analysis, however a number of key features ensure greater consensus & collaboration among users and control over research data.

- **Open-source and locally deployable:** LaMa is open-source and can be deployed locally under organizational infrastructure preventing outside access while allowing collaborative labelling by researchers from an organization. This can benefit from security measures already in place at an organizational level, allow more control over research data, and reduce the possibility of data leakage. Furthermore, due to its open-source nature, the tool can be adapted based on specialized needs.

- **Artifact-based labelling:** LaMa uses an artifact-based approach for labelling. An artifact is a short text that contains one key message and can potentially be labeled with one label. If the labeler thinks the corresponding artifact contains multiple messages, he/she can split the artifact accordingly during the labelling process.

- **Collaborative labelling:** With LaMa multiple researchers can simultaneously label the same set of text artifacts. During labelling, newly created labels are immediately shared with other labelers, which facilitates the reuse of existing labels. Furthermore, a LaMa project can be configured so that it requires one artifact to be labeled by more than one labeler to reduce individual bias.

- **Multi labelling:** LaMa allows each artifact to be labelled with multiple labels. This feature is particularly useful if a researcher wants to label an artifact from more then one viewpoints. These viewpoints, which are called label types, can be configured during project creation.

- **Conflict resolution:** To ensure consensus during collaboration, automatic conflict detection has been implemented. A conflict occurs when one artifact is labeled differently by

---

[4]https://labelstud.io/

multiple labelers. LaMa users can view these disagreements, which facilitates a dialog among corresponding labelers to agree on a label. Having this dialog and resolving the disagreement are very important for reducing individual bias during the labelling process.

- **Themes:** LaMa allows users to group labels into themes, thereby providing help in the classification and the analysis of the data. Furthermore, themes can be categorized hierarchically further aiding in the analysis and classification process.

- **Traceability:** LaMa keeps a record of all the changes made to the artifacts, labels and themes. These changes are visible on the details page of the corresponding entities. This adds an extra layer of traceability.

Conclusion and future work

LaMa is an open-source web-application for thematic labelling of qualitative data. Its key-assets are obtaining insight into data by hierarchically grouping them into labels & themes and facilitating better collaboration between users through features such as collaborative labelling & conflict resolution.

LaMa has been extremely beneficial for analyzing the data of an ongoing semi-structured interview research. We hope that LaMa can aid future qualitative research by the features it provides and by being extendible due to its open-source nature. Meanwhile, we plan to extend this tool by creating a more intuitive user experience and adding features such as document-based labelling, AI-based assistance, and labelling of audio, video & images.

Acknowledgements

References

Aghajani, E. (2022). Labeling machine: A light-weight web application for researchers to label their data with ease. In *GitHub repository*. GitHub. https://github.com/emadpres/labeling-machine

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101. https://doi.org/10.1191/1478088706qp063oa

Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE guide no. 131. *Medical Teacher*, *42*(8), 846–854. https://doi.org/10.1080/0142159X.2020.1755030

Muctadir, H. M. (2022). Labeling machine: A light-weight web application for researchers to label their data with ease. In *GitHub repository*. GitHub. https://github.com/muctadir/labeling-machine

Pope, C., Ziebland, S., & Mays, N. (2000). Analysing qualitative data. *British Medical Journal (BMJ)*, *320*(7227), 114–116. https://doi.org/10.1136/bmj.320.7227.114

Richards, K. A. R., & Hemphill, M. A. (2017). A practical guide to collaborative qualitative data analysis. *Journal of Teaching in Physical Education*, *37*(2), 225–231. https://doi.org/10.1123/jtpe.2017-0084