

Unit 3 - Research Proposal Review:

Which of the methods described in this week's reading would you think would suit your purpose?

QuestionPro (2021) as well as the University of Liverpool Academic Skills (n. d.) differentiate between qualitative and quantitative research methods. In their definition, qualitative research is used to conduct a general exploration of a research question, without a specific research question in mind beforehand. Through this process of exploration, questions worth answering are then brought up. A different approach is followed by quantitative research. Here, a clearly defined hypothesis is evaluated and a specific research question is aimed to be answered. As I focus on using machine learning algorithms for sentiment analysis, generally a quantitative research suits my approach better. Especially, a sub-group of quantitative research that the University of Liverpool Academic Skills (n. d.) names “relationship-based” will fit well to my planned research. In this sub-group, research is conducted to understand the relationship between variables. In my case this will relate to answering the question which factors (e.g., words or emojis) influence the sentiment of a statement (Grover and Banati, 2022; Van Hee et al., 2018).

Furthermore, QuestionPro (2021) differentiate between exploratory and conclusive research. While exploratory research aims at generating general insights about a topic, conclusive research follows a specific research purpose, collects data for this purpose, follows a formal approach in analysing this data and then answering the research question based on this analysis. Conclusive research is either descriptive or causal. Causal research fits my purpose best as it aims to find causal relationships between variables. This goes one step further than finding correlations by not only

finding that two variables develop in the same direction but by specifying that the change of one variable is the reason for the change of the other variable (Rohrer and Murayama, 2023; Docquier et al., 2024). This sort of finding is vital in answering my research question and thus this approach best fits to my purpose. However, there are many possible ways in which causal research can be conducted. Sage (2021), in a visualisation, give some examples. Among these are different types of cluster analysis and discriminate analysis. However, in my specified research question I already partly define which causal method I will follow. I aim at classifying or clustering statements and assign them to be related to a positive, neutral, or negative sentiment.

Concluding, within quantitative methods I will conduct a causal analysis to get a maximum of information on the cause-and-effect-relationship between different variables. In the causal analysis, I will use several different neural network architectures to assess the sentiment of wordings in different domain contexts. Comparing these models and their results, I will derive a model that presents a best fit to the research question. I will see which wordings and contexts have which effect on the sentiment, deriving causality.

Which data collection methods would you consider using?

To build a sentiment analysis system, I need a vast amount of data that covers not only sentences and messages made available by individuals, but also a clear mapping to a sentiment. I will not be able to measure any of this data. Thus, I search for different data sets. Firstly, I search for a data set that holds data on general conversations (Kazanova, 2013). This is used to train the first model, which is rather general. Secondly, I search for a data set from a different domain to apply domain specification from the original model on a different data set. I could, for example use a data set that shows IMDB film reviews for this (Kurmi, 2019). Testing the model across different domains and for these two different data sets improves its generalisability, which is an important research objective. Additionally, I must also ensure that I test all trained models within the research, i.e., the optimised model as well as all models used as benchmarks, are trained and evaluated on the same data sets to ensure comparability of the achieved results (Willmington et al., 2022; Ashbaugh and Zhang, 2024).

Which required skills will you need to have or develop for the chosen project?

My project will require significant skills in Python, which I will use as programming language to train the neural networks needed for my analysis. Furthermore, I will also require knowledge about how to best create a neural network architecture for the given task. While I am acquainted to programming in Python and using its libraries for creating and training neural networks, I only have little experience in training neural networks. Up to this point, I mainly trained convolutional neural networks for recognising objects and images and building image classifiers. While this work gives me some confidence in putting together a neural network and fine-tuning it to improve its performance, this given use case where I will put together a neural network to analyse text data and use advanced natural language processing algorithms I have not worked with yet.

However, while this newly acquired skill is necessary it should not be too challenging or overwhelming as there are plenty of examples in the literature how such a model could look like. Examples are given by Liu et al. (2019) and He et al. (2021) which I can use and then only need to fine-tune for my specific purposes. Scanning the literature and then improving the model by critically evaluating the applied architectures and using best practices for building neural networks to improve them will be a good starting point for the task. I expect to improve my skillset in this area significantly throughout working on this task and this way additionally profit from it.

List of References:

Ashbaugh, L. and Zhang, Y. (2024) 'A Comparative Study of Sentiment Analysis on Customer Reviews Using Machine Learning and Deep Learning', *Computers* 13(12).

Available at: <https://www.mdpi.com/2073-431X/13/12/340>

Docquier, D., Di Capua, G., Donner, R. V., Pires, C. A. L., Simon, A. and Vannitsem, S. (2024) 'A comparison of two causal methods in the context of climate analyses', *Nonlinear Processes in Geophysics* 31(1), pp. 115-136. Available at:

<https://npg.copernicus.org/articles/31/115/2024/>

Grover, V. and Banati, H. (2022) 'Understanding the Sarcastic Nature of Emojis with SarcOji', *Proceedings of the Fifth International Workshop on Emoji Understanding and Applications in Social Media*, pp. 29-39. Available at:

<https://aclanthology.org/2022.emoji-1.4/>

He, P., Liu, X. Gao, J. and Chen, W. (2021) 'DeBERTa: Decoding-enhanced BERT with Disentangled Attention', *arXiv preprint*. Available at:

<https://arxiv.org/abs/2006.03654>

Kazanova, M. M. (2013) Sentiment140 dataset with 1.6 million tweets. Available at:

<https://www.kaggle.com/datasets/kazanova/sentiment140> (Accessed: 06 April 2025)

Kurmi, A. (2019) IMDB-sentiment-analysis. Available at:

<https://github.com/Ankit152/IMDB-sentiment-analysis> (Accessed: 06 April 2025)

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M.,

Zettlemoyer, L. and Stoyanov, V. (2019) 'RoBERTa: A Robustly Optimized BERT

Pretraining Approach', *arXiv preprint*. Available at:

<https://www.semanticscholar.org/paper/RoBERTa%3A-A-Robustly-Optimized-BERT-Pretraining-Liu-Ott/077f8329a7b6fa3b7c877a57b81eb6c18b5f87de>

QuestionPro (2021) What is research? Available at:

<https://www.questionpro.com/blog/what-is-research/> (Accessed: 09 February 2025)

Rohrer, J. M. and Murayama, K. (2023) 'These Are Not the Effects You Are Looking for: Causality and the Within-/Between-Persons Distinction in Longitudinal Data

Analysis', *Advances in Methods and Practices in Psychological Science* 6(1).

Available at: <https://journals.sagepub.com/doi/10.1177/25152459221140842>

Sage (2021) Sage Research Methods: Methods Map. Available at:

<https://methods.sagepub.com/methods-map/> (Accessed: 09 February 2025)

University of Liverpool Academic Skills. (n.d.) Introduction to research methods and methodologies. Available at: <https://www.youtube.com/watch?v=nv7MOoHMM2k>

(Accessed: 09 February 2025)

Van Hee, C., Lefever, E. and Hoste, V. (2018) 'SemEval-2018 Task 3: Irony Detection in English Tweets', *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 39-50. Available at:
https://www.researchgate.net/publication/325447252_SemEval-2018_Task_3_Irony_Detection_in_English_Tweets

Willmington, C., Belardi, P., Murante, A. M. and Vainieri, M. (2022) 'The contribution of benchmarking to quality improvement in healthcare. A systematic literature review', *BMC Health Services Research* 22(1). Available at:
<https://pmc.ncbi.nlm.nih.gov/articles/PMC8812166/>