IAW DISCUSSION PAPERS

iaw@iaw.edu | www.iaw.edu

IAW Diskussionspapiere Nr. 143 November 2023 | No. 143 November 2023

Content Analysis with Language Models – Exploring a Zero-Shot Learning Approach

Philipp Kugler
Yvette Bodry
René Kalweit
Andreas Koch
Marcel Reiner
Tobias Scheu

Institut für Angewandte Wirtschaftsforschung e.V.
Schaffhausenstraße 73 | 72072 Tübingen | Germany
Tel.: +49 7071 98960 | iaw@iaw.edu

ISSN: 1617-5654

IAW-Diskussionspapiere

Das Institut für Angewandte Wirtschaftsforschung (IAW) Tübingen ist ein unabhängiges außeruniversitäres Forschungsinstitut, das am 17. Juli 1957 auf Initiative von Professor Dr. Hans Peter gegründet wurde. Es hat die Aufgabe, Forschungsergebnisse aus dem Gebiet der Wirtschafts- und Sozialwissenschaften auf Fragen der Wirtschaft anzuwenden. Die Tätigkeit des Instituts konzentriert sich auf empirische Wirtschaftsforschung und Politikberatung.

Dieses IAW-Diskussionspapier können Sie auch von unserer IAW-Website als pdf-Datei herunterladen:

https://www.iaw.edu/iaw-diskussionspapiere.html

ISSN 1617-5654

Weitere Publikationen des IAW:

- IAW-News (erscheinen 4x jährlich)
- IAW-Impulse
- IAW-Kurzberichte
- IAW Policy Reports
- IAW-Forschungsberichte

Möchten Sie regelmäßig eine unserer Publikationen erhalten, dann wenden Sie sich bitte an:

IAW Tübingen e.V.
Schaffhausenstraße 73, 72072 Tübingen
Telefon: 07071 98960
iaw@iaw.edu

Aktuelle Informationen finden Sie auch im Internet unter:

https://www.iaw.edu/

Content Analysis with Language Models – Exploring a Zero-Shot Learning Approach

Philipp Kugler* (IAW Tübingen)
Yvette Bodry (Universität Hohenheim)
René Kalweit (IAW Tübingen)
Andreas Koch (IAW Tübingen)
Marcel Reiner (IAW Tübingen)
Tobias Scheu (IAW Tübingen)

Abstract

In this paper, we discuss whether and how Natural Language Processing (NLP) can be integrated into the workflows of conventional qualitative research to support the researcher. We focus on textual data which the researcher wants to analyse using predefined categories. We explore the possibility of carrying out the coding step using NLP instead of assigning codes manually. Integrating such models into the qualitative research process makes the research more accessible and comprehensible for third parties and can contribute to moving research more into the direction of open science. Our study indicates that the procedure is potentially able to identify the core results. However, the findings also indicate that there are weaknesses which strongly depend on the specific text and the research question under investigation. We find that off-the-shelf language models are not able to distinguish between related topics as clearly as humans can. Moreover, humans are able to understand and classify passages which only implicitly refer to a topic. Findings suggest that off-the-shelf language models frequently fail to identify such passages. We conclude that an important step to reliably apply language models in qualitative research consists of improving the models. Nevertheless, in their current state, off-the shelf language models can be used to validate the results obtained by manual coding. This could make qualitative research more traceable and fully reproducible.

*Corresponding author: Schaffhausenstr. 73, D-72072 Tübingen; philipp.kugler@iaw.edu.

We would like to thank Jens Vogelgesang and all the participants of the ZEW Workshop on "Big Data and new developments in research data centers 2022" and THE Workshop 2022 for their comments. Elisa Bührle is acknowledged for excellent research assistance.

1 Introduction

In recent years, methods in Natural Language Processing (NLP), a field of artificial intelligence developing statistical methods that enable computers to process and analyse vast amounts of text data, have significantly improved. In this context, Chat-GPT, a chatbot created by the company OpenAI, is the most prominent example. These advances raise the question to what extent such methods can support qualitative research.

Generally, integrating such methods into qualitative research is not straightforward, since qualitative methods consist of a wide range of different concepts and types of data, ranging from text to image and video data (Hammersley 2013). In this paper, we focus on data that is available in written form (especially interview data, but also written documents in general), that the researcher wants to analyse using predefined categories (also known as deductive coding). One popular approach in this context is content analysis (Mayring 2015, Silverman 2015), which is an important element of the mixed methods designs in order to analyse research topics in a comprehensive and differentiated manner (e.g., Groenewoud et al. 2022, Lindsäter et al. 2023, Valzadeh et al. 2020).

In this paper, we pursue two objectives. First, we explore a workflow in which NLP methods can be integrated into conventional content analysis by taking over isolated tasks that can be monitored and controlled by the researcher. Second, we discuss whether the approach is viable in practise and provide some directions for future research to improve it and make it more reliable.

Previous research on integrating NLP methods into qualitative research has focused on statistical models such as Topic Modelling (e.g., Baumer et al. 2017, Leeson et al. 2019) or Bag-of-Words models (e.g., Isoaho et al. 2021) to extract features from textual data. Topic Modelling summarises large amounts of texts into frequently occurring themes that are clusters of expressions or words. Using this technique, the researcher cannot influence the topic delimitations as the procedure is *unsupervised*. Moreover, since only the prevalence of words is considered, information about the context and semantic meaning of words is lost. The Bag-of-Words model counts the words of a predefined vocabulary in a document, but still faces identical shortcomings. Mikolov et al. (2013) developed the neural network Word2Vec which allows to estimate semantically meaningful representation of individual words by incorporating their surrounding context. To this end, Word2Vec represents each word as a vector of real numbers, where geometric proximity corresponds to semantic similarity. Thus, compared to previous models, Word2Vec captures a meaningful representation of language as similar words are close to each other in a vector space. Besides, the vectorisation of words ensures efficient machine processing and allows the application of linear algebra to examine the relationship between words. The idea of vector representations of words has been developed into more advanced and powerful language representation models (Vaswani et al. 2017). As one of the most influential downstream models upon them, BERT (Bidirectional Encoder Representation from Transformers) performs complex tasks, such as identifying multiple connotations of language.

Our paper is most closely related to Leeson et al. (2019), which is a proof of concept paper to evaluate the potential of NLP to analyze qualitative data. They conclude that there are two potential ways to incorporate NLP methods: either, after coding to check the accuracy of codes, or before coding to guide the creation of the codebook. In this paper, we go one step further. We analyse whether we can carry out the coding step using NLP instead of assigning codes manually. To this end, our analysis is based on Sentence BERT (SBERT), a model that represents entire sentences in a vector space. Compared to fully unsupervised methods such as Topic Modelling, the procedure incorporates all the researcher's knowledge about the content of the text. By imposing constraints and defining topics, a researcher narrows down the search space (Yu et al. 2011).

Similar to Leeson et al. (2019), our approach seeks to improve two aspects of qualitative research. First, conventional qualitative research is time-consuming and labour-intensive (Winter 2000). We argue that the approach is less time intensive and can help to analyse contextualized, unstructured data at lower costs, especially when the sample size becomes larger (Abram et al. 2020, Clifford & Marcus 1986). Second, to evaluate the quality of qualitative research, there exist various criteria of goodness concerning the collection and analysis of data as well as reporting of the empirical findings. In qualitative research, criteria of goodness are not universal, but a matter of discourse related to epistemological schools of thought and derived assumptions (Kuckartz 2016). In general, many of these criteria are desirable but at the same time difficult to realise (O'Connor & Joffe 2020, Strübing et al. 2018). NLP methods, such as BERT, can facilitate the justification of criteria in an improved manner. In particular, the intercoder reliability (O'Connor & Joffe 2020, MacPhail et al. 2016, Feng 2015, Campbell et al. 2013) and intracoder reliability (O'Connor & Joffe 2020) are two core criteria of goodness that are often considered problematic. This may raise concerns about the reproducibility and consistency of the qualitative approach (Abram et al. 2020, Yu et al. 2011). In our approach, the coding process is precisely documented, making the results more traceable and fully reproducible. Hence, it is easier to credibly comply with the criteria of goodness that are directly connected with coding data, namely the intercoder and intracoder reliability.

In order to assess the viability and quality of the automated encodings generated by SBERT, we use empirical material from a recent study assessing the reactions and strategies of firms and employees to the introduction and the subsequent increase of the statutory minimum wage in Germany. Our analysis draws on one topic discussed in these interviews, namely how the introduction and subsequent increases of the minimum wage affects the wage structure of the firms. This question was recently analysed by Koch et al. (2020) using manual coding. Therefore, we are able to compare (1) how the manual codes deviate from the codes created by SBERT and (2) how the final results derived by manual coding differs from the results derived by SBERT.

Based on our experience, we believe that outsourcing the coding step has numerous benefits and considerable potential. Although the researcher has to invest some time to get used to the handling of language models, the coding step with NLP can be less time consuming, especially for a considerable number of observations and vast amounts of text. More importantly, the coding step is well documented and easy to replicate, which moves qualitative research more into the direction of open science (e.g., Branney et al. 2023, Class et al. 2021, Humphreys et al. 2021). However, future research has to improve and adjust the language models to the specific topic of the data. Our analysis indicates that currently, the model struggles to distinguish between related topics as clearly as humans do. Moreover, humans understand and differently classify passages which only implicitly mention a topic of interest. We find that SBERT often is not able to identify such passages. Both *weaknesses* undoubtedly directly affect the quality of the coding. Whether the interpretation and conclusions differ depends on the exact texts and research question. In our specific case, the results derived by manual coding and those derived by SBERT do not substantially differ. In our opinion, this is impressive given that we use a general version of SBERT off-the-shelf without fine tuning it to the specifics of the data, which is usually done (e.g., Lee et al. 2020; Sun et al. 2019).

The paper is structured as follows. Section 2 describes the procedure. In section 3, we conduct a case study to analyse whether automated encodings using SBERT are viable in practise. Finally, section 4 discusses the results and outlines some perspectives for future research.

2 Analysing text data with known topics

The goal of this paper is to discuss whether and how NLP methods can be incorporated into qualitative content analysis when the researcher has knowledge about the topics that are prevalent in the text

and intends to encode them using a predefined code system. We propose to apply a procedure that directly fits into conventional qualitative workstreams. Thereby, the coding step is carried out with NLP methods.

First, we divide the text into small parts, for example into sentences. Second, we encode each sentence of the text as well as the predefined topics of the code system into a numerical vector to ensure machine readability. Third, we assign a predefined topic to each sentence exploiting the properties of machine-readable encoding. Finally, we can proceed as usual with analysing the encoded sentences. In the following, we explain how the sentences are encoded (section 2.1) and how codes are assigned to these sentences (section 2.2). Finally, we summarise the procedure (section 2.3).

2.1 Representing text with BERT

Generally, computers or statistical models are not able to read text in the same sense humans do. Therefore, it is imperative in NLP to explore ways for presenting text in a machine-readable format. Recently developed models are built on the idea of the distributional hypothesis from linguistics (Jurafsky & Martin 2021). The hypothesis states that words which share similar contexts tend to have similar meanings. For example, synonyms such as “car” and “automobile” are likely to have common surrounding word environments. Therefore, state of the art models represent text as numerical vectors which encode the contextual surrounding of the word. Such vectors are called embeddings. For example, the popular neural network model Word2Vec (Mikolov et al. 2013) represents a single word as a vector of probabilities. The model predicts the likelihood for every word in a defined vocabulary to occur in the surrounding context of the word of interest. Comparing word embeddings with each other cannot only provide information on the similarity or dissimilarity of the words’ meaning, but the difference of word embeddings is itself a vector which can have semantic meaning. The resulting geometric space of adding the embeddings of the words “London” and “Germany” and subtracting “UK” is close to the representation of “Berlin”.

For an embedding to be meaningful, the model has to take into account that words can have diverse semantic meanings. For example, in the sentences “I **left** the event at the river **bank**” and “I turned **left** into the **bank** to get cash”, the meanings of the words “left” and “bank” depend on their context. Whereas humans can easily identify the meaning of the words, a statistical model needs to follow a defined pattern. Early models were only able to *read* the sentence from one direction. However, if the words were mapped to their embeddings from left to right, the encoding model would not know which semantic meaning of the word “bank”, i.e., “bank/waterside” or “bank/cash-machine”, would be correctly meant. The language representation model BERT (Bidirectional Encoder Representations from Transformers) is able to encode a word by jointly depending on the left and right context (Devlin et al. 2018). Its improved performance compared to other models on a variety of different NLP tasks, such as answering questions and translation, has stirred up the machine learning community.

When analysing textual data, we are often not interested in the meaning of a single word, but in a set of connected words such as a sentence or a series of sentences. A naïve approach would be to represent each sentence by pooling (for example averaging) the word embedding of all words in a sentence. However, this does not work well in practice (Reimers and Gurevych 2019). Therefore, we rely on Sentence-BERT (SBERT, Reimers and Gurevych 2019), a model that fine-tunes the pooled embeddings obtained from BERT to increase semantic richness.[1] Further, SBERT is computationally rapid and cheap.

[1] To be precise, we used Cross English & German RoBERTa fine-tuned from Philip May and open-sourced by T-Systems-onsite: https://huggingface.co/T-Systems-onsite/cross-en-de-roberta-sentence-transformer.

2.2 Encoding text without labels

After the embeddings of each sentence are computed by SBERT, we want to assign each sentence a predefined code. We follow recent work that shows that language models can handle tasks for which they were not explicitly trained for (see for example Radford et al. 2019). The approach can be considered as a zero-shot learning approach, i.e., an approach that is able to recognize new concepts by simply describing them (Romera-Paredes and Torr 2015). In our case, this means that we want to predict the code of a sentence only based on the description of that predefined codes. Since the description of the codes consist of pure text, we can proceed in an identical way as for the sentences and embed both sentences and codes into the same vector space using SBERT. Therefore, in this step, the predefined codes need to be operationalised by describing the topic they are referring to. Afterwards, SBERT is used to compute the embedding of the description. We assign a code to a sentence when both are semantically close to each other and therefore likely to concern the same topic. As described above, comparing the embeddings with each other provides information on the similarity or dissimilarity of the sentences' meaning. Therefore, we compute the cosine similarity between the embedding of the code and the embedding of the individual sentences. The cosine similarity is a measure of the similarity of two vectors that depends on the angle between them. The cosine defines a range between -1, indicating the opposite meaning and 1, indicating the same meaning of the two sentences. A predefined code is assigned to a sentence if they are similar in terms of the cosine similarity. Algorithm 1 summarizes this procedure.

Algorithm 1: Analysing text data with known topics

Algorithm
1: Partition the texts into small parts. It is convenient to use sentences, since the input of SBERT is limited to 512 tokens, which roughly correspond to words.
2: Compute the embedding of each part.
3: Operationalise the topics, i.e., the conventional codes the text is about and compute the embedding of each topic/code.
4: For each topic/code, compute the cosine similarity for each sentence. Assign the code to those sentences that are similar given a chosen minimum threshold similarity.
5: Analyse the coded text in the conventional manner.

2.3 Practical issues

The procedure raises some issues that the researcher has to address. First, the method heavily relies on the operationalisation of codes. If the quality of the operationalisation is poor, the assignment process will also be poor. If operationalization is too vague, then assigned codes will be too imprecise. Moreover, the procedure is difficult to apply when codes are semantically similar. There are no clear rules and specification on how topics must be defined and how semantically different topics should be framed in order to be distinguished. To obtain a well-defined operationalisation, the researcher has to become familiar with the ways language models practically work. We therefore recommend experimenting with the model before starting the analysis. Second, the procedure analyses each sentence

independently. However, the context, i.e., the surrounding sentences, is usually important to completely understand the sentence. Thus, in our application, we do not only assign codes to a single sentence, but also to the five sentences preceding and following the encoded sentence. Third, the similarity between code and sentence is represented by the cosine similarity, i.e., a continuous number. The lower the number, the greater the dissimilarity between the code and the text partition. However, a binary decision is needed to code the sentence (similar or not similar). The decision, which numerical value still indicates similarity is an individual decision, the researcher has to make since there is no general cut-off threshold. Rather, it depends on the specific topic and text.

3 Are automated encodings using SBERT viable in practice?

In general, the performance of our procedure is difficult to measure since the process of coding is always subjective and there are no objective encodings.[2] Nevertheless, we need to obtain a better understanding of how the zero-shot learning approach described in section 2 works and whether it is viable in practice. In order to assess the integrity and quality of the automated encodings generated by SBERT, we compare the results derived from this approach with those derived from a conventional approach (i.e., manual coding). To this end, we use a selection of our material to analyse (1) how the manual encodings differ from the encodings made by SBERT, and (2) how the final results derived by manual encoding differ from the results derived by SBERT encodings.

More precisely, we use empirical material from a recent study assessing the patterns of reaction and strategies of firms and employees in the context of the introduction and the subsequent increase of the statutory minimum wage in Germany (Koch et al. 2020). The following analysis is based on 60 literally transcribed guided interviews with firms' management or staff departments. They have been conducted by telephone or personally between May and September 2019. On average, the interviews took 44 minutes and contain 8,400 spoken words. The interviews are based on a comprehensive interview guideline comprising various topics concerning the reactions and the behaviour of firms to the minimum wage as well as firms' strategies to cope with the minimum wage. In order to identify and clarify the commonalities and the differences between the results of the (already existing) manual coding and the automated coding by SBERT, we focus on one specific topic of the interviews – namely the effects of the minimum wage on the internal wage structure of firms.

3.1 The minimum wage and internal wage structures

From the perspective of an affected firm[3], both the introduction as well as every subsequent increase of the minimum wage causes at least an increase of the wages of the employees directly affected by the minimum wage. Moreover, the overall wage structure of a firm employing both affected and non-affected workers changes. If the firm does nothing else than complying to the legal requirements, the wage differences between affected workers (rising wages) and non-affected workers (unaltered wages) diminish. Furthermore, all employees who previously earned wages below the new minimum wage level will now earn at least the minimum wage. As a result, any former wage differentiations between these employees may cease. To analyse how firms cope with these changes in the overall

[2] In purely qualitative research, the intercoder reliability serves as a measure or process to ensure comparability of results between different individuals coding the same text (Campbell et al., 2013; MacPhail et al., 2016; O'Connor & Joffe, 2020).

[3] We call firms „affected by the minimum wage" if they have at least one employee with a wage below the minimum wage at the time of its introduction or the subsequent increases.

wage structure, Koch et al. (2020) develop a mainly deductive coding scheme complemented by selective inductive elements (as it is usually done when conducting a content analysis). Independently from these coding schemes, we operationalise each reference in the guideline as an input to SBERT. Table 1 summarises the relation between the guideline, coding scheme by Koch et al. (2020) and the defined topics for SBERT. While topics in SBERT almost entirely consist of buzzwords, the coding scheme used by Koch et al. (2020) is more detailed, contains further explanations and includes more context information. These differences in coding schemes make the differences in their operation very clear. SBERT solely relies on a measure of semantic similarity. In the course of conducting the study, we found that further information and clarification introduce disturbance in the embeddings. In contrast, further information and clarification is beneficial, or even necessary for humans to decide whether and how text passages should be encoded.

Table 1: Operationalization of the coding scheme

Topic in Guideline	Topic in SBERT	Coding scheme in Koch et al. (2020)
Development of the wage sum and affected wage groups	Minimum wage adjustment, wage increases, wage or salary of employees above the minimum wage	Immediate change in wage costs or wages due to the introduction or increase of the minimum wage (so no wage reactions here), amount of wage difference (bite) (hourly wage before/after increase/introduction of minimum wage), number (or proportion of the total workforce) of employees directly affected by the introduction or increase in the minimum wage
Changes of employees not directly affected by the minimum wage, overall internal wage structure, gaps between wage groups	Wage hierarchies and inner-company wage structure, distances between wage groups	Pay gaps between individual departments, functional and job levels, gender, (age) cohorts, etc., entire company wage structure, wage cuts for employees with hourly wages above the minimum wage, measures to restore wage gaps
Changes regarding the documentation and recording of working hours Effort for the implementation of the documentation and recording of working hours	Duty of documentation and recording of working time, electronic cash register systems, payroll accounting	Additional effort due to the documentation requirement (hours per week / month / year) coping with the additional effort (more staff, external consultants, digitization of the recording, etc.)

3.2 Comparison with SBERT

3.2.1 Comparing the encodings

First, we compare the SBERT encodings with the human encodings by Koch et al. (2020). The results of this comparison are presented in Figure 1. Each row depicts one interview. Coded areas are marked by colours and the colour shows the origin of the encoding. Yellow bars indicate human encodings. Depending on the similarity measures, encodings from SBERT are either blue, red or purple. Blue areas show SBERT encodings which are highly similar to the operationalization (cosine similarity of ≥ 0.6) and the red colour indicates encodings that are quite similar to the operationalization (cosine similarity

between 0.55 and 0.59). Areas encoded in purple indicate that texts are more distant to the operationalization, but still meaningful (cosine similarity between 0.49 and 0.54). Hatched areas mark an overlapping between a SBERT and human encoding.

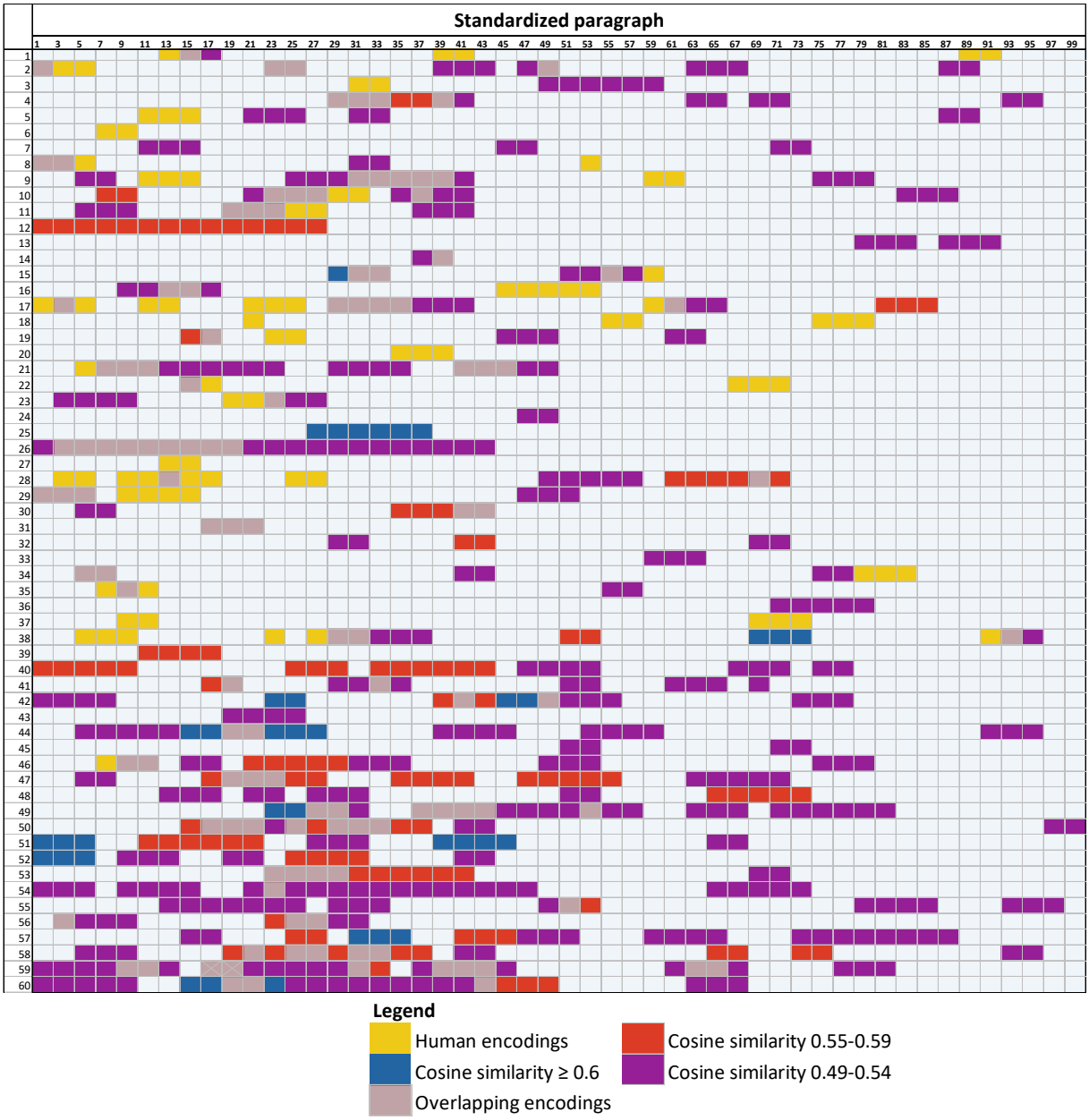
The results show some overlaps between SBERT and human encodings (hatched areas). However, quite often text passages do not overlap at all (coloured areas). The number of SBERT encodings not corresponding to a human encoding depends on the exact threshold. Mostly, these are passages that are rather distant to the operationalization (i.e., purple), e.g.: *"In the end, the employee certainly benefits from the fact that he has more in his hands with the rising minimum wage, be it twelve euros or whatever. [...] In the end, the whole thing is financed by the worker himself, in that he has higher expenses for his daily life, so that everyone else around him can also pay the increased wages, which have to be maintained through hierarchies and distances. In principle, we are only turning the inflation screw. Nothing more"* (interview 4). In some cases, encodings are very close to the operationalization in terms of the cosine distance. It turns out that these text passages discuss topics such as work organization, working time or even general assessments of the minimum wage. For example, SBERT assigns a code with a high proximity to the operationalisation, even though there is no specific connection to a wage hierarchy in terms of content: *"Let's say, helpers or assistants who help me a lot before Christmas are for the most part from my circle of friends and acquaintances, of course not so much changes, because the wage is secondary, yes? And from that point of view, of course it's a matter of whether I've done an hour more or an hour less at Christmas, it doesn't really matter*" (Interview 51).

There is no doubt that these topics (or words used when someone talks about these topics) have some semantic relation with the topic of adjusting wage structures. However, in terms of content, these topics crucially differ from each other. For example, SBERT coded the following quote from a small farm in the context of wage hierarchies, even though it is about holiday entitlements of temporary workers as a consequence of the minimum wage law: *„Well, they get the 10 euros for the hour and of course they get all these holiday entitlements. Of course, this is calculated much more accurately than before. Before, of course, someone who, I know, had only been a student for two months or half a year, who had helped out somewhere, had no holiday entitlement calculated. [...] Of course, you have to calculate exactly how much holiday entitlement he would have had during that time."* (Interview 58).

Moreover, there are human encodings without any overlapping SBERT encoding. It is noticeable that SBERT has not assigned an encoding to passages in which wage structures are described more implicitly, without relying on vocabulary that specifically points to such a topic. Thus, the following passage from a call centre was coded by a human, but not by SBERT: *"What we had to do when we introduced the minimum wage, we had to make much more of a one-size-fits-all. [...] What we had to do was to distribute it differently, so that employees who were not so good and more or less took it easy earned more than before and others who had earned a lot before due to their good performance had to cut their bonuses"* (Interview 20). This quote illustrates changes in the wage structure without explicitly naming them as such and therefore couldn't be detected by SBERT.

In summary, the results show that zero-shot learning is able to detect similar encodings compared to humans. However, off-the shelf language models like, non-fine-tuned SBERT, have considerable weaknesses. First, they are not able to distinguish between crucially different topics that rely on words that are semantically quite close. Second, the language model is not able to detect text passages where the topics are discussed more implicitly.

Figure 1: Comparison between manual and automated encodings



3.2.2 Comparing the findings

In this section, we compare the final results obtained by manual coding described and analysed in Koch et al. (2020) with those obtained by encoding via SBERT.

As wage differences between employees reflect differences between the employees' qualifications, occupations, tasks, competencies and personal characteristics, a modification of the firm's internal wage structures (i.e., the distances between different wage groups) might be undesirable and disadvantageous in the view of the firm. In this sense, Koch et al. (2020) find that firms which adjust wages exclusively for workers affected by the minimum wage, report that it has become more difficult (i.e., more expensive) to reflect the performance and the qualifications of employees in their compensation. Other firms with a similar behaviour report declining job satisfaction, especially of higher qualified employees, a decline in team spirit and, ultimately, a deterioration in the working atmosphere. Such firms fear an increasing exit of skilled workers that do not benefit from the minimum wage. Other firms state that issues also emerged with regard to differentiations between marginally employed workers

and workers subject to social insurance contributions, as the former usually do neither pay any taxes nor contributions to social insurance and thus frequently have significantly higher net wage gains than the other group (cf. Koch et al. 2020, p. 56f).

If, on the other hand, firms try to keep wage structures unaltered, they either face increasing personnel costs due to the wage adjustments in the higher wage groups, or they have to take organisational measures or precautions in order to be more efficient and to thereby avoid or compensate the rising wage costs. For instance, firms report that poorly qualified workers with low wages even have been dismissed due to the huge lack of the quality of their work in comparison to other (low wage) workers. Some firms adjusted the wages of employees, not directly affected by the minimum wage, gradually or with a delay. Other firms report that they changed their whole pay scheme, e.g., by strengthening performance-related compensation (cf. Koch et al. 2020, p. 59ff).

Although encodings differ between SBERT and those made by Koch et al. (2020), the analysis leads to the same conclusions. In our case, there is a clear intersection between the findings of both, SBERT and manual encoding. This includes especially the core findings by Koch et al. (2020) concerning diminishing illustration of qualification levels in earnings, negative effects on general working atmosphere in the firms and declining job satisfaction of workers that earn more than the minimum wage. Therefore, we can conclude that SBERT identifies the same core results as Koch et al. (2020).

However, there are some results that deviate. Using SBERT, we do not find results that are only marginally related to the change in wage structures (e.g., the dismissal of workers according to their productivity). Moreover, as indicated above, SBERT relates some findings to wage structures that deviate from the core understanding of the specific topic. We also obtain findings that relate to the general increase of wages or special payments.

In summary, the comparison identifies two issues. First, language models cannot distinguish between related topics as clearly as humans. This is because language models are only based on the semantic distance, which does not necessarily intersect with a human's core understanding and definition of a topic. Second, humans, in contrast to language models like SBERT, understand and classify passages which only implicitly mention some topic. These *weaknesses* undoubtedly directly affect the quality of the coding. Whether the interpretation and conclusions differ depends on the exact text and research question. In general, only the latter issue leads to an incomplete picture, since some relevant passages might not be encoded. The first issue only leads to an excess amount of encoded passages which are not relevant for interpretation and have to be ignored or "decoded" by the researcher. The second issue might lead to wrong conclusions because of *sloppy* coding. In our application, the differences in SBERT encoding and manual encoding are rather large, whereas the first issue is particularly prevalent. However, through "decoding", the interpretation and conclusions reached from the relevant encodings do not crucially differ from the results elaborated by Koch et al. (2020).

4 Discussion and perspectives

This paper explores and discusses if and how statistical language models can be used for coding text data. The method directly fits into the conventional procedures of content analysis since it replaces the coding step usually executed manually by a researcher with a language model. Applying the procedure enables researchers to analyse large amounts of text data without losing the in-depth view, which distinguishes qualitative research from other disciplines. Moreover, the coding step can be fully understood and replicated by a third party, which is a crucial criterion of goodness (i.e., inter- and intracoder reliability). Therefore, we think that integrating language models into the qualitative research process makes the research more accessible and comprehensible for third parties and can help to move research more into the direction of open science.

Our study indicates that the procedure is potentially able to identify the core results. However, the results also show that there are weaknesses that strongly depend on the exact text and research question. We find that off-the-shelf language models cannot differ between related topics as clearly as humans. This is because language models are only based on the semantic distance, which does not necessarily intersect with a human's core understanding and definition of a topic. Moreover, humans understand and classify passages which only implicitly mention some topic. We find that SBERT often is not able to identify such passages. This weakness poses a great risk of overlooking relevant text passages.

In general, these weaknesses can be partly addressed by choosing a well-considered operationalisation of topics. However, our study indicates that this might not be enough. One important step to reliably apply language models in qualitative research consists of improving the models. In this paper, we only use an off-the-shelf version of SBERT which was trained on many different texts – given that, the results are impressive in our opinion. When analysing specific data, one might account for the nature of it. In our case, we might account for slang, incomplete sentences, idioms and, most importantly, the special vocabulary that often occurs in the context of labour market research. For example, BioBert (Lee et al. 2020) was developed to analyse biomedical text data. Training such a model goes beyond the scope of this explorative paper, but should be added to the future research agenda on integrating NLP methods into qualitative research. Nonetheless, we conclude, that in their current state off-the-shelf language models can be utilized to validate the results obtained by manual coding. This could make qualitative research more traceable and fully reproducible.

References

Abram, M. D., Mancini, K. T., & Parker, R. D. (2020). Methods to integrate natural language processing into qualitative research. *International Journal of Qualitative Methods*, *19*, 1609406920984608.

Baumer, E. P., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, *68*(6), 1397-1410.

Branney, P. E., Brooks, J., Kilby, L., Newman, K., Norris, E., Pownall, M., Talbot, C. V., Treharne, G. J. & Whitaker, C. M. (2023). Three steps to open science for qualitative research in psychology. *Social and Personality Psychology Compass, 17(4)*, e12728.

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42*(3), 294-320.

Class, B., de Bruyne, M., Wuillemin, C., Donzé, D., & Claivaz, J. B. (2021). Towards open science for the qualitative researcher: From a positivist to an open interpretation. International Journal of Qualitative Methods, 20, 16094069211034641.

Clifford, J., & Marcus, G. E. (Eds.). (1986). *Writing culture: the poetics and politics of ethnography: a School of American Research advanced seminar*. University of California Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feng, G. C. (2015). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *11*(1), 13.

Groenewoud, A. S., Leijten, E., van den Oever, S., van Sommeren, J., & Boer, T. A. (2022). The ethics of euthanasia in dementia: A qualitative content analysis of case summaries (2012–2020). Journal of the American Geriatrics Society, 70(6), 1704-1716.

Hammersley, M. (2013). Defining qualitative research. *In What is Qualitative Research?* (The 'What is?' Research Methods Series, pp. 1-20). London: Bloomsbury Academic.

Humphreys, L., Lewis Jr, N. A., Sender, K., & Won, A. S. (2021). Integrating qualitative methods and open science: Five principles for more trustworthy research. Journal of Communication, 71(5), 855-874.

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*, *49*(1), 300-324.

Jurafsky, D., & Martin, J. H. (2021). Vector Semantics and Embeddings. In *Speech and language processing* (pp. 96-126), Vol. 3.

Koch, A., Kirchmann, A., Reiner, M., Scheu T., Zühlke, A. & Bonin, H. (2020). *Verhaltensmuster von Betrieben und Beschäftigten im Kontext des gesetzlichen Mindestlohns*. IZA Research Report No. 97, Bonn.

Kuckartz, U. (2016*). Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung* (3rd ed.). Beltz Verlag, Weinheim Basel.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Leeson, W., Resnick, A., Alexander, D., & Rovers, J. (2019). Natural Language Processing (NLP) in qualitative public health research: a proof of concept study*. International Journal of Qualitative Methods*, *18*, 1609406919887021.

Lindsäter, E., Svärdman, F., Rosquist, P., Wallert, J., Ivanova, E., Lekander, M., ... & Rück, C. (2023). Characterization of exhaustion disorder and identification of outcomes that matter to patients: Qualitative content analysis of a Swedish national online survey. Stress and Health.

MacPhail, C., Khoza, N., Abler, L., & Ranganathan, M. (2016). Process guidelines for establishing intercoder reliability in qualitative studies. *Qualitative Research*, *16*(2), 198-212.

Mayring, P. (2015). *Qualitative Inhaltsanalyse* (12th ed.). Beltz Verlag, Weinheim Basel.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: debates and practical guidelines. *International Journal of Qualitative Methods*, *19*, 1609406919899220.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Romera-Paredes, B., & Torr, P. (2015, June). An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning* (pp. 2152-2161). PMLR.

Silverman, D. (2015). *Interpreting qualitative data*. Sage.

Strübing, J., Hirschauer, S., Ayaß, R., Krähnke, U., & Scheffer, T. (2018). Gütekriterien qualitativer Sozialforschung. Ein Diskussionsanstoß. Zeitschrift für Soziologie, 47(2), 83-100.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18 (pp. 194-206). Springer International Publishing.

Valizadeh, L., Zamanzadeh, V., Ghahramanian, A., Musavi, S., Akbarbegloo, M., & Chou, F. Y. (2020). Adolescent cancer survivors' experiences of supportive care needs: A qualitative content analysis. Nursing & health sciences, 22(2), 212-219.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Winter, S. (2000, May 15). Quantitative vs. Qualitative Methoden. *PKR Universität Karlsruhe*.

Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. The Qualitative Report, 16(3), 730-744.

Institut für Angewandte Wirtschaftsforschung e.V.
an der Universität Tübingen

Schaffhausenstraße 73
72072 Tübingen
Telefon: 07071 98960
iaw@iaw.edu
www.iaw.edu

IAW Institut für Angewandte Wirtschaftsforschung e.V.

an der Universität Tübingen