Automatic Detection of Everyday Social Behaviours and Environments from Verbatim Transcripts of Daily Conversations

Kristina Y. Yordanova[1], Burcu Demiray[2,3], Matthias R. Mehl[4], Mike Martin[2,3]
1 *Department of Computer Science*, *University of Rostock*, Rostock, Germany
2 *Department of Psychology*, *University of Zurich*, Zurich, Switzerland
3 *University Research Priority Program "Dynamics of Healthy Aging"*, *University of Zurich*, Zurich, Switzerland
4 *Department of Psychology*, *University of Arizona*, Tucson, AZ, USA
kristina.yordanova@uni-rostock.de, b.demiray@psychologie.uzh.ch, mehl@email.arizona.edu, m.martin@psychologie.uzh.ch

***Abstract*—Coding in social sciences is a process that involves the categorisation of qualitative or quantitative data in order to facilitate further analysis. Coding is usually a manual process that involves a lot of effort and time to produce codes with high validity and interrater reliability. Although automated methods for quantitative data analysis are largely used in social sciences, there are only a few attempts at automatically or semi-automatically coding the data collected in qualitative studies. To address this problem, in this work we propose an approach for automated coding of social behaviours and environments based on verbatim transcriptions of everyday conversations. To evaluate the approach, we analysed the transcripts from three datasets containing recordings of everyday conversations from: (1) young healthy adults (German transcriptions), (2) elderly healthy adults (German transcriptions), and (3) young healthy adults (English transcriptions). The results show that it is possible to automatically code the social behaviours and environments based on verbatim transcripts of the recorded conversations. This could reduce the time and effort researchers need to assign accurate codes to transcribed conversations.**

***Index Terms*—social behaviour analysis, natural language processing, automated coding**

I. Introduction and Motivation

Coding in social sciences is an analytical process, in which data in both qualitative or quantitative form are analysed and categorised in order to facilitate further analysis [25]. This process is usually conducted manually, requires the involvement of at least two annotators to validate the codes and is very time consuming and error prone process, especially when large quantities of data are collected [7]. The process is even further complicated by the need of producing annotation with high interrater reliability, which in itself includes training the coders for a given problem [32]. As the size of the data typically collected in studies increases [5], it becomes difficult, even impossible, to code the data manually.

To cope with the increase of data, crowdsourcing has sometimes been used to produce the codes [13]. Although it provides an interesting strategy to coping with large amounts of data, it still suffers from the bias of the participating annotators, especially when they have different social backgrounds and thus different interpretation of the data [33].

Although the use of automated methods for codes analysis is widely spread in social sciences [5], the automated coding is an emerging research field. What is more, it is usually conducted by computer scientists without much involvement of the domain experts in the development of the codes [5].

To address the above problems we propose an automated approach for identifying social behaviours and environments from verbatim transcripts of daily conversations. The transcripts originate from 30-second audio recordings collected in the course of the participants' daily lives. As there are already variety of methods and commercial tools for automatic transcription of speech [35][1], we go one step further and look at the ability of automated approaches to identify the psychological codes associated with these transcripts based on data annotated by domain experts. The targeted codes themselves are identified by expert psychologists researching healthy ageing in our society, while the features used for assigning the codes are developed jointly by psychologists and computer scientists, all authors in this work.

The contribution of the paper is twofold: 1) we propose a procedure for automatically identifying social behaviours and interactions in transcripts of everyday conversations; 2) we investigate a new application domain to text analysis and classification by testing the methodology on three real world datasets from the psychology domain.

The paper is structured as follows. Section II presents the related work in automated coding from textual sources. Section III presents our approach to automatically coding verbatim transcripts. In Section IV, we present the evaluation methodology and the corresponding results. Finally, Section

K. Yordanova is funded by the German Research Foundation, grant number YO 226/1-1; at the time this research was conducted, K. Yordanova and M. R. Mehl were funded by the University of Zurich's Digital Society Initiative in the context of the DSI Fellowships and Collegium Helveticum.

[1]We have to note that automated methods for audio transcribing in social sciences are in their infancy as state of the art tools have shown poor quality at least when using the Electronically Activated Recorder [18] to record the data. Nevertheless, here we assume that such tools have potential to providing accurate transcripts.

978-1-5386-9148-9/19/$31.00 ©2019 IEEE

V discusses the proposed approach. The work concludes with short discussion about future work in Section VI.

II. Related Work

Behaviour analysis is a well-known topic in the pervasive computing community. There are numerous works addressing the recognition of persons' activities, situation, or environment [4], [22], [29], [31]. These are, however, usually based on sensor data, the targeted behaviours are identified by computer scientists and usually address physical behaviour. There are also some attempts to recognise the person's behaviour based on the combination of textual descriptions and sensor observations [15], [30]. These works also address physical behaviour and do not build on domain expertise for identifying psychological variables associated with this behaviour.

Apart from identifying physical behaviour, there is a lot of research on recognising the person's affect [23]. It is especially centred on image and speech techniques for emotion recognition [1], [10], but there is also a growing amount of work analysing textual sources, such as tweets, in order to identify the affect of the person or the crowd [6]. A popular approach in classifying tweets is using neural networks and deep learning to obtain the tweet's affect [11], [24]. For such approaches to perform, however, the model requires large quantities of annotated data, which is not always available in social sciences studies. What is more, these works are centred on the person's affect and do not address other psychological factors that could be of interest for the social sciences[2].

To support the coders in social sciences, there are some attempts at providing automatic suggestions of labels based on some previous examples. For example, in [7] the authors apply a bag of words approach combined with part of speech information and information about the person's location to automatically identify codes of interest. As expected, they report on poor results in the cases where there are not enough samples of a given class. The unbalanced class distribution is a problem observed in our datasets, as well.

Another work for semi-automated coding relies on manually coding small portion of the data by domain experts, then crowdsourcing a larger portion to be labelled by non-experts following the rules created by the experts [12]. These labelled data are later used to train convolutional neural network that is able to code a much larger dataset. Although it provides an interesting solution to the problem of producing sufficient quantities of labelled data, this approach has questionable annotation quality typical for untrained coders [2].

Yet another work attempting to support the coding in social sciences proposes the training of a support vector machine with a labelled dataset, then asking coders to manually evaluate and correct the labels [28]. This second step is necessary due to the unbalanced class distribution, where there are a lot of negative samples, but very few positive examples. In our work, we propose an alternative solution to coping with unbalanced data by applying data augmentation techniques.

Apart from supervised methods for transcript analysis, there are works that rely on unsupervised methods to analyse the textual data. For example, in [9] the authors use topic models to analyse psychotherapy and medication therapy transcripts. They, then, manually analyse the resulting topic clusters, concluding that they contain clinically relevant content. In our work, we also rely on topic models. We, however, use the resulting topics to generate features that are later used for training a classifier to automatically code transcripts.

Apart from the problems associated with unbalanced classes in the data, a commonly observed problem is that most works on automated coding in social sciences rely on state-of-the-art machine learning tools [7], [12], [28]. This is similar to the above discussed problem of computer scientists developing models for automated behaviour analysis without understanding of the domain-specific factors associated with the behaviour. In this case, however, social scientists use standard models that are often not tuned for the underlying data. This highlights the need of experts from both computer science and social sciences to work together in developing appropriate models for automated coding.

III. Methods and Materials

In this section, we present our approach to assigning psychological codes from the verbatim transcripts. The novelty here is the data analysis process as a whole, which results from the specific application domain. Although the separate parts in the proposed procedure have been applied to different machine learning (ML) problems, to our knowledge this is the first attempt at practically investigating a workflow for "easy-to-reproduce" automatic coding of socio-psychological variables. What is more, the proposed process is a synergy of expertise from both social and computer sciences, to adjust the data analysis workflow based on the codes identified by domain experts. In that sense, the procedure and the empirical evaluation open a new application domain to ML methods.

A. Procedure for Automated Coding

To manually code the transcripts, an annotator assigns a code from a given category to each transcribed conversation (which we call a sample). Table I shows example conversations. As psychologists are interested in different categories of socio-psychological variables, they consider each category separately and assign one of the variables (which we call codes or classes) in a category to each sample. They then repeat the process for all categories. The list of categories and the codes they contain can be seen in Tables II and III. For example, if they are evaluating the first conversation in Table I for the category "mood", they will assign one of the following classes "laugh", "sing", "cry", "mad", "sigh", or "none". Then they will continue with the next category, e.g. "self function" and assign one of its classes (codes) to the conversation. The categories are disjunct, in other words,

[2]This could partially be explained with the effort needed to annotate sufficient amount of data with detailed and reliable codes in various categories as opposed to coding just the affect.

TABLE I
EXAMPLE TRANSCRIPTS OF SOCIAL CONVERSATIONS.

Transcripts
Hey, I'm about to leave. Okay. Do you have your key to get, um never mind I found it.
Are we going, I like the black one better. I like that red one.
Yes, um I heard about Saturday. I didn't get back there till late.
What happened? Um, there's still one picture on this camera. I don't know, I don't know how to take it.
It's like class on Friday after the exam. So I'm determined I'm probably not going to have to take the final. It's different.

the codes of one category do not appear in another and the categories are independent of each other.

Our goal is to automate this process. We consider each of the categories as a separate classification problem, where the classifier assigns one of the codes in the category to each conversation, then repeats the procedure with the next category. Similar to the manual coding, if we want to find out what is the mood in the first conversation in Table I, the classifier will classify the text as one of the following classes "laugh", "sing", "cry", "mad", "sigh", or "none". It will then consider the next category, e.g. "self function" as a separate classification problem and it will classify the same text as one of the following classes "explaining oneself", "evaluating oneself", "reassuring ones beliefs" or "none of the above".

To be able to classify a conversation, the transcribed conversations first have to be processed. Fig. 1 illustrates the procedure for a given category C. We begin with all

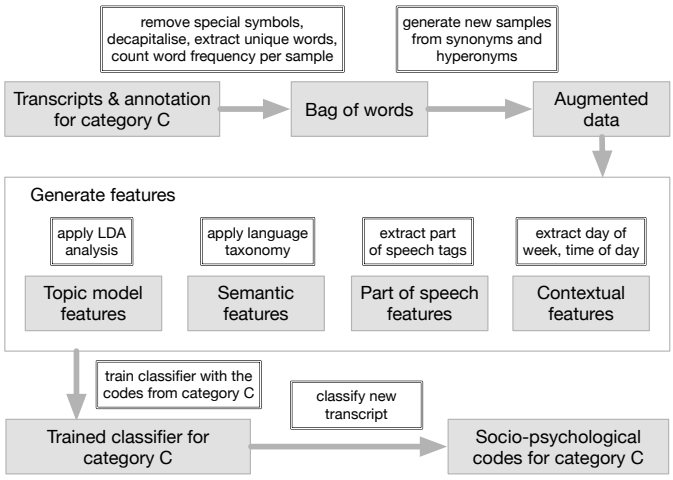

Fig. 1. The proposed process for detection of social behaviours and environments from transcripts.

transcripts and the annotation for category C. The annotation here is the code from C assigned to each transcript. We first convert the transcripts into bag of words. This is done by removing any special symbols in the transcripts, decapitalising all words, extracting all unique words in the whole corpus with conversations, and finally counting the number of times each unique word in the corpus appears in each of the transcripts. In that manner, each transcript is converted into a sample in vector representation with m attributes, where m is the number of unique words in the corpus and each attribute represents the number of times a word appeared in the given conversation (see Section III-B). As the classes in the data are unbalanced (see Section III-C), we perform data augmentation to generate new samples from existing ones. We do that by replacing words from the original sample with their synonyms or hyperonyms. Using the data in this format for classification is impractical as we have as many attributes as the unique words in the corpus and many of them do not contain useful information. For that reason, we attempt to reduce the number of attributes by extracting different statistical, lexical, semantic, and contextual features from the bag of words. We use the following strategies: 1) we group the attributes into different topics with the help of latent Dirichlet allocation (see Section III-D); 2) we replace the words with their semantic abstractions with the help of language taxonomy (see Section III-F); 3) we replace the words with their part of speech meaning (see Section III-E); 4) we replace the words with different contextual features such as time of day, day of the week, etc. (see Section III-G). We use the extracted features (or subsets of the features) to train state-of-the-art classifiers such as decision tree, random forest, and support vector machine. The classifier then assigns one of the classes (codes) from category C to a new transcript.

B. Data Preparation

The first step is the data preparation. We start by generating bag of words from the original transcripts. This is also known as a vector space model, where for each document in a corpus, the frequency of appearing words is represented as a vector. We first extract all unique words in the dataset. We, then, remove any non-alphabetical or numerical symbols, and any stop words and we decapitalise the words. In difference to typical bag of words approaches, we do not perform stemming or lemmatisation, as we assume that the form of the word contains relevant lexical information, such as temporal focus or number of participants in conversation (singular / plural).

Then, for each transcribed audio sample, we count how many times we have observed a given word. This produces a matrix $B = \mathbb{R}^{n \times m}$, where n is the number of samples (or transcripts) and m is the number of unique words in the corpus (or attributes). A sample here is a transcript from a single audio file. The resulting matrix B is the input for the rest of the transformations, described in this section.

C. Coping with Unbalanced Classes

Initial analysis of the data showed that the classes in each of the categories are unbalanced. With small exceptions, there was a large number of samples for one class, and very few for some of the remaining classes. For example, for the category "conversation type" (see Table II for more details on the categories), we have 159 samples for the class "small talk", 2156 for the class "substantive talk", 28 for "personal disclosure", 470 for "practical talk", 17 for "gossip" and 386 for "none". Even more unbalanced is the category "ageing" with 2 samples

in class "age related talk", 9 in "memory related" and 3206 in "none". To cope with this problem, we utilise the idea of data augmentation. Data augmentation is typically used in image recognition, where to cope with unbalanced classes, new data are generated by performing transformations on the original data [20]. As we are dealing with textual data, to transform a sample we utilise a language taxonomy. We used the taxonomy of English language WordNet [21]. From the taxonomy, for each word in a sample we extract its synonyms or hyperonyms in case no synonyms were found. We replace some of the words in the sample with the corresponding synonyms, creating a derivative of the original sample. The idea of using synonyms to simulate textual data has already been explored in works such as [34]. As words having exactly or nearly the same meanings are relatively few, synonym-based augmentation can be applied to only a small percentage of the vocabulary [14]. As we believe that it is improbable for different people to use the same or almost the same words to express a given social behaviour, in cases where no synonym is available, we use the hyperonym of the word. In other words, we take the word that has a "type-of" relation with the original word. For example, the word "Alice" has the hyperonym "person", the word "box" has the hyperonym "container".

For a bag of words B with m unique words and a dictionary $D = (W_B, W_{syn})$, which consists of m tuples of a unique word in a bag w_B and its synonym w_{syn}, we generate new samples by using the following procedure. If a class c in a category has less samples than a given threshold Th, then we generate new samples by p times transforming the samples in class c and adding them to the existing samples where $p = f_{max}/f_c$. Here, f_{max} is the number of samples in the class with the most samples, and f_c is the number of samples in class c. We use a sliding window of q words over the original W_B. Each time we transform the samples from class c, we change q out of m original words with their synonyms, where $q = m/p$, with m being the number of unique words in the bag. That way we ensure that each transformed sample is different from the original sample and from the rest of its derivations. Each transformed sample receives the class of the original sample. We then set the frequency of the original word in a sample to 0 and transfer the original frequency to the synonym word. The new bag of words B_{new} then contains $W_{B_{new}}$ unique words where $W_{B_{new}} = W_B \cup W_{syn}$. The matrix $B_{new} = \mathbb{R}^{n_{new} \times m_{new}}$ where $m_{new} = \textit{length}(W_{B_{new}})$ and $n_{new} = n + \sum_c p_c \times f_c$ with p_c being the number of times a sample is augmented in class c. In that manner we produce new samples that make the data more uniformly distributed among the classes. We repeat the procedure for each category, as the different categories are uniquely imbalanced.

Initial evaluation showed that using the samples from the bag of words directly as input for a classifier does not produce satisfactory results. We also attempted using the data directly for training a neural network (NN), however the NN achieved accuracy of only about 30% to 40%. We concluded that the amount of data is not enough to train a well performing NN and instead concentrated on performing additional transformations of the data before we used them for classification.

D. Topic Models for Feature Extraction

We performed principal component analysis (PCA) to remove some of the words in the bag and reduce the number of attributes we will use for classification. It showed that only a small percentage of the data can be explained with the first three components (about 0.4% of the words are explained in the first component, 0.3% in the second, and 0.2% in the third, and the remaining components are relatively uniformly distributed). In other words, we were unable to reasonably remove words. To cope with this problem, we applied latent Dirichlet allocation (LDA), which is a type of topic model. LDA is a generative statistical model that groups sets of observations into unobserved groups [3]. Each group contains observations with some similarity between them. In our case, the observations are the vectors describing how often a word was observed in a sample. We attempted to find $k < m$ topics that group the m words in the bag of words. As we are able to generate arbitrary number of topics $k \leq m$ out of the m unique words, one problem is identifying appropriate k. We follow the approach proposed in [8], which uses Gibbs sampling to determine the probability of words in a corpus W given k topics, $P(W|k)$. By varying k, the goal is to identify the value of k that produces the model with the highest likelihood.

After performing LDA, we are interested in the estimated per-topic word probabilities $\varphi_{z_{i,j}}$. $\varphi_{z_{i,j}}$ tells us with what probability a word w appears in a topic z. We use it to reduce the number of variables in B by transforming B from $n \times m$ matrix to $n \times k$ matrix $T = \mathbb{R}^{n \times k}$ where k is the number of topics. For each $t_{jl} \in T$, we compute its value by

$$t_{jl} = \sum_{i=1}^{m} b_{ji} * \varphi_{z_{ij}}, \tag{1}$$

where $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$, and $l \in \{1, \ldots, k\}$.

E. Using Linguistic Information

As we are interested in identifying the temporal focus of a conversation, codes such as giving or receiving information / advice, or if the person is talking to one or multiple persons, we assume that the part of speech (POS), to which a word belongs, will give us additional useful information. For example, the word "girl" is a noun in its singular form, indicating one person, while the word "responded" is a verb in its past form, indicating that the focus is the past.

We parse all words in the bag to obtain the POS tag (i.e. part of speech). We transform the bag of words B from $n \times m$ matrix to $n \times p$ matrix $Q = \mathbb{R}^{n \times p}$, where p is the number of unique POS tags and each $q_{jl} \in Q$ represents the number of words from the original sample assigned to this tag.

$$q_{jl} = \sum_{i=1}^{m} b_{ji} * \gamma_{w_i}, \textit{ with } \gamma_{w_i} = \begin{cases} 1, & \text{if } w_i \text{ has tag } g_l, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $i \in \{1, \ldots, m\}$ with m being the number of words, $j \in \{1, \ldots, n\}$ with n the number of documents, $l \in \{1, \ldots, p\}$ with p the number of tags, and g_l is the tag at position l.

F. Using Semantic Information

As we explained, one strategy to reducing the number of words in the data is using topic models. Another strategy is to use semantic similarities. That is, we no longer look into the similarities in the word vectors, but we try to abstract the semantic meaning of the words and thus to reduce the number of attributes. To achieve that, we once again employ a language taxonomy. Our aim is to identify $r < m$ abstract concepts that describe the m unique words in our corpus. Language taxonomies such as WordNet create a concepts hierarchy expressed through its hyperonyms. This is the structure we want to use in order to reduce the number of attributes.

To extract the concepts hierarchy, we start with the set of words W. For each word $w \in W$ we recursively search for its hyperonyms. This results in a hierarchy where the bottommost level consists of the elements in W and the uppermost level contains the most abstract word, that is the least common parent of all $w \in W$. The least common parent *lcp(a,b)* of two words a and b is the parent on the highest level in the taxonomy that contains both a and b as children.

We transform B from $n \times m$ matrix to $n \times h_u$ matrix $R = \mathbb{R}^{n \times h_u}$, where h_u is the number of unique words on an abstraction level u. To calculate the new feature r, we follow similar procedure as in Formula 2.

$$r_{jl} = \sum_{i=1}^{m} b_{ji} * \lambda_{w_i}, \text{ with } \lambda_{w_i} = \begin{cases} 1, & \text{if } w_i \text{ is child of } c_l, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $i \in \{1, \ldots, m\}$ with m being the number of words, $j \in \{1, \ldots, n\}$ with n the number of documents, $l \in \{1, \ldots, h_u\}$, and c_l is the word at position l on level u of the hierarchy.

G. Using Contextual Information

We also extract some contextual information based on the time when the audio files were recorded. From the time stamps, we are able to obtain information on the day of the week (Monday to Sunday), whether it is weekend or not, and the time of the day (morning, noon, afternoon, evening, or night). We extract this information, as we assume that depending on the day of the week and the time of the day, people have different contexts and thus different social behaviours. As suggested by the domain experts (who are authors in the paper), we also added the number of words in a sample as an additional feature. Although the number of words in a sample is not really contextual feature, for simplicity we include it into the contextual features. This feature was extracted based on the assumption that whether a given person talks more or less also affects their social behaviour.

IV. Evaluation

A. Experimental Setup

To test our approach, we applied it to three datasets containing transcripts of everyday social conversations. Two of the datasets were recorded in Switzerland and are in German and one was recorded in the US and is in English. Example transcriptions can be seen in Table I.

1) Dataset 1: The dataset contains daily conversations of young adults. The study is recorded with the Electronically Activated Recorder [18], an audio recorder that unobtrusively tracks real-world behaviours by periodically recording snippets of ambient sounds while participants go about their daily lives. The study consisted of sixty-one undergraduates from Switzerland (23 males, 38 females, average age 25), who wore the EAR for 4 days. The EAR recorded 30 s of sounds randomly throughout the day, for about 72 times a day, providing 18,039 waking recordings (M = 295 per participant). For each recording, whenever a conversation was recorded, it was transcribed. This resulted in 3,217 samples where conversation was observed. Participants spoke Swiss German. Based on the transcripts, the coders identified the person the participant is talking to, the functions of the conversation, the type of the conversation, the temporal focus of the conversation, the location where the conversation was conducted, the associated activity, and the person's mood. Table II shows the codes associated with each of those categories. Apart from the codes in Table II, an additional "none of the above" class was added to each category to describe samples that do not contain any of the targeted codes. All three datasets were manually coded by two persons with interrater reliability above 80% (which indicates almost perfect overlapping). We did not use self-annotation techniques as they tend to produce low interrater reliability compared to annotation by experts [33].

TABLE II
THE CATEGORIES AND THE CORRESPONDING CODES FOR THE FIRST AND THE SECOND DATASETS, CONTAINING CONVERSATIONS OF YOUNG (1ST DATASET) AND ELDERLY HEALTHY ADULTS (2ND DATASET) IN GERMAN.

Category	Codes
(1) talking to (preson)	self, partner / significant other, daughter / son, kids, other relative, friend / acquaintance, familiar person, stranger, pet, unknown
(2) activity	socialising, intoxicated, working, housework, hygiene, eat / drink, TV, exercise / walk, in transit, sleep
(3) mood	laugh, sing, cry, mad, sigh
(4) self function	explaining oneself, evaluating oneself, reassuring ones beliefs
(5) give advice	teaching, giving advice
(6) receive advice	receiving teaching, receiving advice
(7) support	give empathy / support, receive empathy / support, connecting / intimacy
(8) conversation	conversation
(9) directive function	problem solving, planning, decision making, making goal / progress
(10) valence	negative / positive valence
(11) ageing	age related, memory related
(12) conversation type	small talk, substantive talk, personal disclosure, practical talk, gossip
(13) temporal focus	personal past, others past, present, personal future, others future, time-independent

After processing the dataset, it resulted in a bag with 6,555 unique words. The words were clustered in 12 topics, as 12 clusters showed optimal results according to the method proposed in [8]. Obtaining part of speech tags resulted in 36 unique POS tags. We used the TreeTagger for German

language[3]. To obtain the abstraction class of the words, we used WordNet [4]. As WordNet is in English, we first translated the words to English, then extracted their abstraction. We used abstraction level 4 (that is, the abstract concept was four levels higher on the abstraction hierarchy than the original word) to produce more abstract concepts and smaller number of unique concepts. This procedure resulted in 28 abstract concepts. Finally, we also extracted the day of the week, weekend, the time of the day, and the frequency of words per sample.

2) Dataset 2: The second dataset aims to analyse the daily activities and conversations of healthy older adults. The study is recorded with the Electronically Activated Recorder [18]. The study consisted of 32 healthy older adults from Switzerland (12 males, 20 females, average age 72), who wore the EAR for 4 days. The EAR recorded 30 s of sounds randomly throughout the day, for about 72 times a day, providing 8,846 waking recordings (M = 276 per participant). For each recording, whenever a conversation was recorded, it was transcribed. This resulted in 1,978 samples where conversation was observed. Participants spoke Swiss German in everyday life. As in Dataset 1, based on the transcripts, the coders identified the person the participant is talking to, the functions of the conversation, the type of the conversation, the temporal focus of the conversation, the location where the conversation was conducted, the associated activity, and the person's mood. The categories and codes in this dataset are the same as those in Dataset 1 (see Table II). We performed the same procedure as with Dataset 1. This resulted in a bag of words with 4,553 unique words, and 32 unique POS tags. We grouped the words in 12 topics for simplicity of the evaluation.

3) Dataset 3: The third dataset aims to analyse the daily social behaviour of happy people [17]. The study is once again recorded with the EAR. The study consisted of seventy-nine undergraduates (32 males, 47 females), who wore the EAR for 4 days [27]. The EAR recorded 30 s of sounds every 12.5 min, providing 23,689 waking recordings (M = 300 per participant). For each recording, whenever a conversation was recorded, it was transcribed. This resulted in 7,753 samples where conversation was observed. Based on the transcripts, the coders identified the person the participant is talking to, the purpose of the conversation, the location where the conversation was conducted, the associated activity, and the person's mood. Table III shows the codes associated with each of those categories. Using the same procedure as with Dataset 1 and 2 resulted in a bag of words with 7,461 words, 22 POS tags, and 10 concepts from WordNet. Here we also group the words into 12 concepts for simplicity as the LDA analysis suggested that 11 to 14 topics would be optimal.

4) Procedure:

a) Simulating new samples: We apply the procedure for data augmentation from Section III in order to produce new samples for classes with small number of samples. We use the procedure for Dataset 1 and 3. Dataset 2 is used as validation dataset, so it stays in its original form. The data augmentation is performed separately for each of the categories in Table II and Table III. This is due to the fact that the different categories have different codes distribution for the same data.

[3]http://www.cis.uni-muenchen.de/s̈chmid/tools/TreeTagger/

[4]https://wordnet.princeton.edu/

TABLE III
THE CATEGORIES AND THE CORRESPONDING CODES FOR THE THIRD DATASET, CONTAINING CONVERSATIONS OF YOUNG HEALTHY ADULTS IN ENGLISH.

Category	Codes
(1) talking to	male(s), female(s), mixed sex, cannot tell
(2) purpose	practical / everyday, school / job, small talk, dialogue / converse, gossip, disclosure, validation / self-assurance, support / caring, conflict
(3) location	apartment, outdoor, in transit, rest / bar / cafe, other public location, unknown
(4) activity	radio, TV, computer, study, work, eat, lecture, sport, entertainment, social, sleep
(5) mood	laugh, sing, cry, mad, sigh

b) Features: We generate four types of features. 1) topic model features (TM): features based on the topics identified through LDA; 2) WordNet features (WN): features based on the abstract concepts in WordNet; 3) part of speech features (POS): features based on the POS tags of the words in the bag; 4) contextual features (CF): day of the week, weekend, time of day, number of words.

c) Classification procedure: We use state-of-the-art classifiers with the extracted features: decision tree (DT), support vector machine (SVM), and random forest (RF). The goal is not to evaluate the classifiers but rather to test whether it is possible to automatically code transcripts of daily conversations based on the proposed pipeline and extracted features. For each category in a dataset, the classifiers assigns one of the codes from this category to a new sample.

Experiment 1: We use the augmented data from Dataset 1 and Dataset 3 to perform 10-fold cross validation using DT, SVM, and RF. We perform the procedure for all categories in the datasets. We also vary the combinations of features to identify their effect on the model performance.

Experiment 2: We use the augmented data from Dataset 1 and Dataset 3. We assume that the augmented data is biased, as it is a derivative of the original data. To evaluate the effect of the simulated data on the results, we divide the data for each category in each dataset into training and test datasets. We remove 5 samples from the original dataset as well as all new samples derived from these samples. In case the class did not have derived samples, we just take the first 100 samples from this class. We use these data as test dataset, while the remaining data are used for training. We repeat the procedure for each category in Dataset 1 and Dataset 3.

Experiment 3: We use Dataset 1 as training and Dataset 2 as test data. More precisely, we use the data and the labels from a given category from Dataset 1 to train the model and then we use the data for the same category from Dataset 2 to test the model. We repeat the procedure for all categories. In that manner, we want to test the ability of the model to generalise data.

TABLE IV
THE RESULTS (IN TERMS OF ACCURACY, PRECISION, AND SPECIFICITY) WHEN APPLYING DECISION TREE (DT), RANDOM FOREST (RD), AND SUPPORT VECTOR MACHINE (SVM) TO THE FIRST DATASET USING ALL FEATURES.

Category	DT			RF			SVM		
	acc	prec	spec	acc	prec	spec	acc	prec	spec
talk to	.73	.70	.97	**.82**	**.81**	**.98**	.68	.65	.96
activity	.70	.67	.97	**.85**	**.85**	**.98**	.65	.64	.96
mood	.64	.60	.91	**.76**	**.65**	**.94**	.65	.61	.91
self fun.	.72	.69	.91	**.74**	.62	**.91**	.73	**.71**	.91
give adv.	.78	.76	.86	**.84**	**.83**	**.90**	.72	.72	.82
rec. adv.	.74	.73	.87	**.94**	**.94**	**.97**	.78	.78	.89
support	.93	.93	.98	**.98**	**.98**	**.99**	.89	.89	.96
conv.	.68	.72	.63	**.74**	**.76**	**.71**	.69	.69	.66
dir. fun.	.87	.88	.97	**.97**	**.97**	**.99**	.89	.90	.97
valence	.97	.97	.97	**.99**	**.99**	**.99**	.99	.99	.99
ageing	.98	.98	.99	**.99**	**.99**	**.99**	.99	.99	.99
conv. t.	.73	.71	.95	**.83**	**.83**	**.97**	.72	.71	.94
tem. foc.	.60	.56	.93	**.78**	**.78**	**.96**	.65	.65	.94

B. Results

1) Experiment 1: We used the extracted features to train three state of the art classifiers (SVM, DT, and RF). We then performed a 10-fold cross validation with the augmented data. The results showed that using the proposed procedure the classifiers were able to recognise the correct code in a given category with accuracy varying (depending on the category) between 60% and 98% for the DT, between 74% and 99% for RF, and between 65% and 99% for SVM. This stands to show that the approach is able to automatically identify relevant codes describing social behaviours and environments. Table IV shows the accuracy, precision, and recall for the DT, SVM, and RF when using all features described in Section III. It can be seen that the RF outperforms the other two classifiers in almost all cases. Performing Wilcoxon test showed that the results of the RF are significantly better than those of the DT and the SVM (p-value of 4.428×10^{-13} between the RF and DT, and 2.228×10^{-11} between the RF and the SVM). On the other hand, the comparison between the SVM and the DT showed no signifiant difference between the results (p-value of 0.68). The variance in the results for the different categories is due to the number of codes in a given category (less codes show better results) and the effect of the data augmentation for very unbalanced classes (classes with few samples produced derivatives that were more similar to the original data than those with more samples, affecting the accuracy of the classifier).

TABLE V
THE RESULTS (IN TERMS OF ACCURACY, PRECISION, AND SPECIFICITY) WHEN APPLYING DECISION TREE (DT), RANDOM FOREST (RD), AND SUPPORT VECTOR MACHINE (SVM) TO THE THIRD DATASET USING ALL FEATURES.

Category	DT			RF			SVM		
	acc	prec	spec	acc	prec	spec	acc	prec	spec
talking to	.40	.35	.83	**.64**	**.59**	**.90**	.41	.35	.83
purpose	.65	.60	.96	**.86**	**.84**	**.98**	.57	.52	.95
location	.66	.61	.94	**.88**	**.86**	**.98**	.49	.46	.91
activity	.63	.55	.96	**.76**	**.75**	**.98**	.46	.42	.95
mood	.82	.81	.96	**.97**	**.97**	**.99**	.72	.71	.94

Similar to Dataset 1, the results from Dataset 3 showed that the random forest had the best performance (see Table V). Depending on the category, the DT had accuracy between 40% and 82%, the RF between 64% and 97%, and the SVM between 41% and 72%. Performing Wilcoxon test showed that the results of the RF are once again significantly better than those of the DT and the SVM (p-value of 3.685×10^{-9} between the RF and DT, and 2.289×10^{-18} between the RF and the SVM). In difference to the first dataset, here the comparison between the SVM and the DT showed that the SVM was significantly worse than the DT (p-value of 1.482×10^{-3}). For that reason, in the rest of the experiments we use the RF classifier to present the results.

We also evaluated the performance of the approach when using different combinations of features. Fig. 2 shows the results from Dataset 1 for different features combinations when using RF. Contextual features and those using WordNet alone have the worst performance, while including the POS features and the topic model features improves the performance. It can also be seen that the best features combination depends on the category. When using only single features, the TM features always perform better than the remaining single features (CF, POS, and WN). The performance of the TM is significantly better than that of the rest of the single features (p-value of 1.612×10^{-22} between TM and WN when performing Wilcoxon test, 1.316×10^{-8} between TM and POS, and 9.009×10^{-17} between TM and CF). On the other hand, combinations of features usually perform better than the single features. The combination of CF, POS, and TM (e.g. in the cases of "activity" and "person") or the combination of CF, POS, and WN (e.g. in the cases of "conversation" and "support") produces slightly better results when not all features are taken into account. Also combinations that do not include CF perform slightly worse than those with CF. In case the goal is to reduce the number of features, the combinations of POS and TM or POS and WN show only slightly worse results. Dataset 3 shows similar tendencies.

2) Experiment 2: To evaluate the effect of the augmented data on the performance, we removed samples and their derivatives from the data and used the remaining data for training the classifier. We tested the trained model on the removed samples. Fig. 4 shows the accuracy, precision, and specificity for Dataset 1 when using RF as classifier and all features. While some categories still have very high performance, others such as "give advice" and "temporal focus" show reduced performance. In other words, for these classes the bias in the augmented data played role in the high accuracy. Similar behaviour is observed in the third dataset, where the performance of the first and the fourth category dropped (see Fig. 5). As the number of samples for training and testing differs in Experiment 1 and Experiment 2, we did not perform statistical tests to see whether the results in Experiment 2 are significantly different from those in Experiment 1.

3) Experiment 3: To test the ability of the proposed approach to generalise on new data, we trained a RF with the

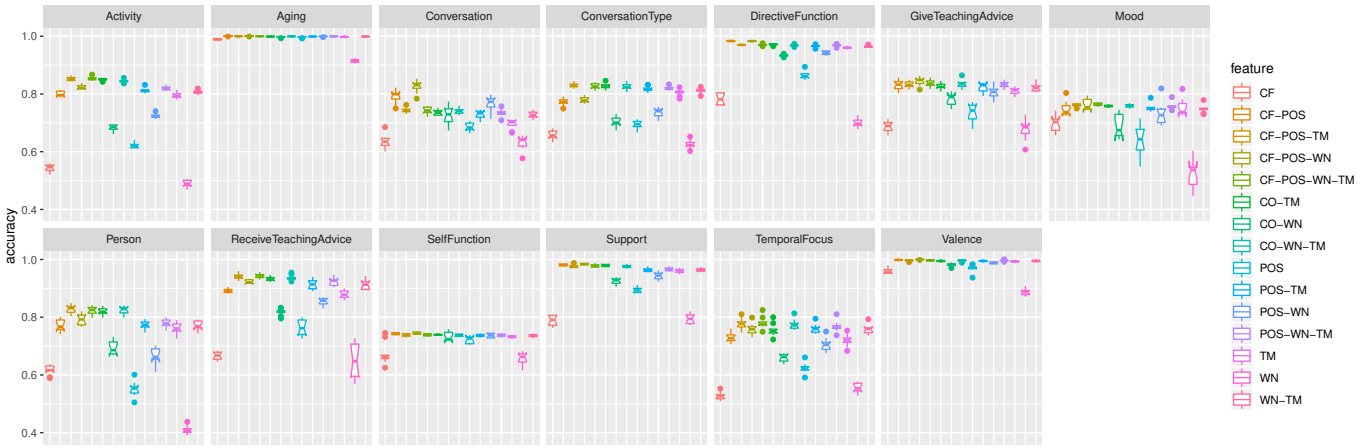

Fig. 2. The results for the targeted categories in Dataset 1 when using different features combinations. CF stands for context features, POS for part of speech features, TM for topic model features, WN for semantic features extracted from WordNet.

Fig. 3. The results for the targeted categories in Dataset 3 when using different features combinations. CF stands for context features, POS for part of speech features, TM for topic model features, WN for semantic features extracted from WordNet.

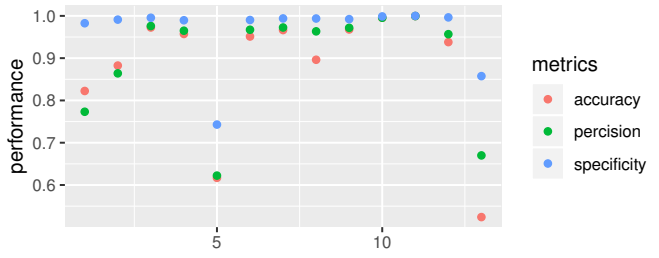

Fig. 4. The results from the first dataset for the targeted categories when removing some samples and all their derivatives from the training data. The numbers in the x axis correspond to the category numbers in Table II.

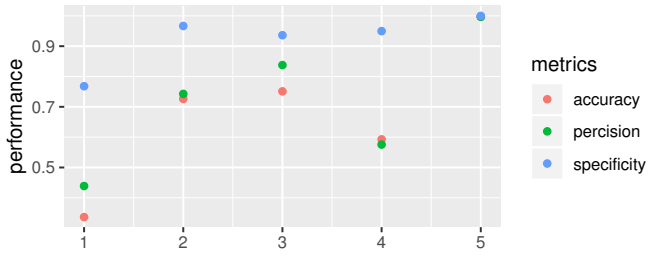

Fig. 5. The results from the third dataset for the targeted categories when removing some samples and all their derivatives from the training data. The numbers in the x axis correspond to the category numbers in Table III.

data from Dataset 1 and then used Dataset 2 for testing. The results showed reduced accuracy for the categories "talking to" (category 1), "activity" (category 2), "conversation type" (category 12), and "temporal focus" (category 13), while the remaining categories had relatively high accuracy (see Fig. 6, the green dots). When looking at the specificity, the opposite effect is observed. The specificity is very low in categories with high accuracy. This is explained with the fact that the classes with high accuracy had high number of negative samples (i.e. "none" class) and very few samples from the remaining classes. The classifier was able to recognise the negative classes but not the few positive samples or there were no positive samples. Fig. 7 illustrates this problem.

We attempted to improve the model performance by adding some samples from Dataset 2 in the training data. More precisely, we took every 3rd sample from the test dataset and added it to the train dataset. We then performed 3-fold cross

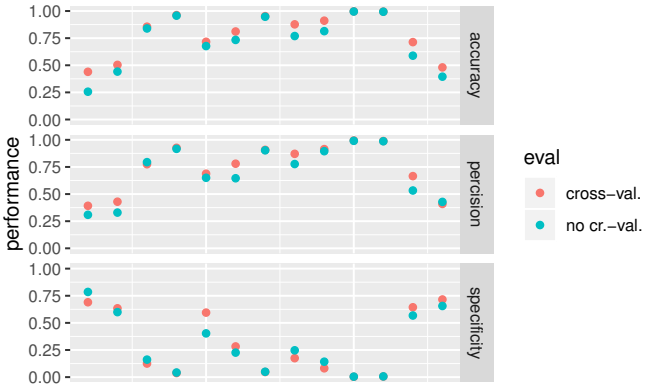

Fig. 6. The results from a RF trained on the first dataset and tested on the second. The red dots show the results when using some of the samples from Dataset 2 for training.

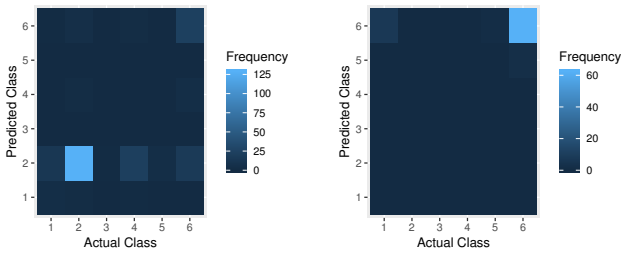

Fig. 7. The confusion matrices for conversation type (left) and mood (right) when using Dataset 1 for training and Dataset 2 for testing.

validation[5]. The results can be seen in Fig. 6 (the red dots). Although there is some slight improvement in some of the classes, generally there was no difference between using samples from Dataset 2 for training, or using only the data from Dataset 1. We performed Wilcoxon test, which confirmed our observations that there is no statistically significant difference between the results when using only Dataset 1 for training and when adding samples from Dataset 2. The reason behind this problem is that there are just not enough samples in some of the classes to contribute for improving the learned model.

V. Discussion

In this work we presented a process for automated coding of social behaviours and environments from verbatim transcripts. We explored different strategies for extracting features from the textual data, including contextual, lexical, and semantic features. We also addressed the problem of augmenting data in order to produce sufficient quantity of training examples. The results showed that our approach is able to recognise the correct labels in different socio-psychological categories with a very high accuracy.

It was also observed that the simulated data produces bias in some of the categories but not in all of them. This could be explained with the way in which the new samples are generated. When many new samples are required, only a few words are exchanged with their synonyms or hyperonyms, which means that the new data are very similar to the original samples. One solution to this problem could be to use a dictionary that contains various synonyms of a given word, then use this richer set of words to replace a larger number of the original words with their substitutes. Another approach could be to perform the data augmentation before creating the bag of words, i.e. on a sentence level.

We observed that when testing the trained model on completely new data, it performs very well in recognising negative samples, but very poorly in correctly classifying the rest of the classes. This problem is due to the fact that there are very few positive examples from a given class. This makes it infeasible to improve the performance even when adding some samples from the new data to retrain the model. One solution here could be to apply data augmentation on the test dataset in order to balance the classes in the data and then to add samples from the augmented data to the training data.

Another approach to improving the quality of the assigned codes could be to manually analyse a small amount of the automatically assigned codes, then to add these (corrected) coded examples to the training dataset and retrain the model. In that manner, a quality control is achieved and in the same time, the effort is still smaller than when a coder has to annotate the whole large dataset manually.

We used LDA for reducing the feature space. Another approach could be using word embeddings such as word2vec [19] or doc2vec [16], which rely on neural networks and a vector representation of the words.

We used a bag of words approach that disregards the sentence structure. The structure of the sentence, however, contains additional semantic information that could be useful for making better label predictions. In other words, apart from extracting the POS-tags, we could also extract the dependencies between words in the transcripts. This approach is similar to the one proposed in [26] for analysing motivational interviews. For extracting these features however, instead of bag of words, the original sentence structure has to be used.

In this work we considered the problem as a single class classification problem. In reality, it might be the case that a sample belongs to more than one class. In that case, the human annotators assigned a "dominant class" to the sample, which we used as the "true class". Another option would be to treat it as a multi-class classification problem.

VI. Conclusion and Future Work

In this work we showed that automated coding approaches could potentially be applied in qualitative studies in social sciences. We plan to use these results as basis for introducing (semi-)automated coding in our future qualitative studies.

Another avenue we plan to pursue, is to forego the transcription of audio data and instead to identify the correct codes as well as additional contextual information based on the audio signal. To achieve that we intend to investigate methods for audio features extraction and deep learning methods.

[5]The reason for using only 3-fold cross validation is that some of the classes had as few as 4 samples.

REFERENCES

[1] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5, Feb 2017.

[2] P. S. Bayerl and K. I. Paul. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Comput. Linguist.*, 37(4):699–725, Dec. 2011.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[4] L. Chen and I. Khalil. Activity recognition: Approaches, practices and trends. In L. Chen, C. D. Nugent, J. Biswas, and J. Hoey, editors, *Activity Recognition in Pervasive Intelligent Environments*, volume 4 of *Atlantis Ambient and Pervasive Intelligence*, pages 1–31. Atlantis Press, 2011.

[5] N.-C. Chen, M. Drouhard, R. Kocielnik, J. Suh, and C. R. Aragon. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Trans. Interact. Intell. Syst.*, 8(2):9:1–9:20, June 2018.

[6] N. Colneri and J. Demsar. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*, pages 1–1, 2018.

[7] K. Crowston, X. Liu, and E. E. Allen. Machine learning and rule-based automated coding of qualitative data. In *Proceedings of ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem*, pages 108:1–108:2, Silver Springs, MD, USA, 2010. American Society for Information Science.

[8] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[9] Z. E. Imel, M. Steyvers, and D. C. Atkins. Computational psychotherapy research: scaling up the evaluation of patient-provider interactions. *Psychotherapy*, 52(1):19–30, 2015.

[10] R. Jiang, A. T. Ho, I. Cheheb, N. Al-Maadeed, S. Al-Maadeed, and A. Bouridane. Emotion recognition from scrambled facial images via many graph embedding. *Pattern Recognition*, 67:245 – 251, 2017.

[11] Z. Jianqiang, G. Xiaolin, and Z. Xuejun. Deep convolution neural networks for twitter sentiment analysis. *IEEE Access*, 6:23253–23260, 2018.

[12] D. Karamshuk, F. Shaw, J. Brownlie, and N. Sastry. Bridging big data and qualitative methods in the social sciences: A case study of twitter responses to high profile deaths by suicide. *Online Social Networks and Media*, 1:33 – 43, 2017.

[13] A. Kittur, J. V. Nickerson, M. Bernstein, E. Gerber, A. Shaw, J. Zimmerman, M. Lease, and J. Horton. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1301–1318, New York, NY, USA, 2013. ACM.

[14] S. Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457. Association for Computational Linguistics, 2018.

[15] T. Kollar, S. Tellex, D. Roy, and N. Roy. Grounding verbs of motion in natural language commands to robots. In O. Khatib, V. Kumar, and G. Sukhatme, editors, *Experimental Robotics*, volume 79 of *Springer Tracts in Advanced Robotics*, pages 31–47. Springer Berlin Heidelberg, 2014.

[16] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196. JMLR.org, 2014.

[17] M. Mehl, S. Vazire, S. E Holleran, and C. Shelby Clark. Eavesdropping on happiness: Well-being is related to having less small talk and more substantive conversations. *Psychological Science*, 21:539–41, 04 2010.

[18] M. R. Mehl, J. W. Pennebaker, D. M. Crow, J. Dabbs, and J. H. Price. The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, 33(4):517–523, Nov 2001.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[20] A. Mikoajczyk and M. Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pages 117–122, May 2018.

[21] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[22] J. Parviainen, J. Bojja, J. Collin, J. Leppnen, and A. Eronen. Adaptive activity and environment recognition for mobile phones. *Sensors*, 14(11):20753–20778, 2014.

[23] S. Poria, E. Cambria, R. Bajpai, and A. Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98 – 125, 2017.

[24] Y. Ren, Y. Zhang, M. Zhang, and D. Ji. Context-sensitive twitter sentiment classification using neural network. In *AAAI Conference on Artificial Intelligence*, 2016.

[25] J. Saldana. *The coding manual for qualitative researchers*, chapter An Introduction to Codes and Coding, pages 1–42. SAGE Publications Ltd, 2016.

[26] M. Tanana, K. Hallgren, Z. Imel, D. Atkins, and V. A. Srikumar. Comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment*, 65:43–50, 2016.

[27] S. Vazire and M. R. Mehl. Knowing me, knowing you: the accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology*, 95(5):1202–1216, Nov 2008.

[28] J. L. S. Yan, N. McCracken, S. Zhou, and K. Crowston. Optimizing features in active machine learning for complex qualitative content analysis. In *Workshop on Language Technologies and Computational Social Science, 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, June 2014.

[29] J. Ye, S. Dobson, and S. McKeever. Review: Situation identification techniques in pervasive computing: A review. *Pervasive Mob. Comput.*, 8(1):36–66, Feb. 2012.

[30] K. Yordanova. From textual instructions to sensor-based recognition of user behaviour. In *Companion Publication of the 21st International Conference on Intelligent User Interfaces*, IUI '16 Companion, pages 67–73, New York, NY, USA, 2016. ACM.

[31] K. Yordanova and T. Kirste. A process for systematic development of symbolic models for activity recognition. *ACM Transactions on Interactive Intelligent Systems*, 5(4):20:1–20:35, December 2015.

[32] K. Yordanova and F. Krüger. Creating and exploring semantic annotation for behaviour analysis. *Sensors*, 18(9):2778:1–2778:22, 2018.

[33] K. Yordanova, A. Paiement, M. Schrder, E. Tonkin, P. Woznowski, C. M. Olsson, J. Rafferty, and T. Sztyler. Challenges in annotation of user data for ubiquitous systems: Results from the 1st arduous workshop. Technical Report arXiv:1803.05843, arXiv preprint, March 2018.

[34] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 649–657, Cambridge, MA, USA, 2015. MIT Press.

[35] K. Ziman, A. C. Heusser, P. C. Fitzpatrick, C. E. Field, and J. R. Manning. Is automatic speech-to-text transcription ready for use in psychological experiments? *Behavior Research Methods*, Apr 2018.