Annotating Patient Information Needs in Online Diabetes Forums

Julia Romberg[1], **Jan Dyczmons**[2,3,4], **Sandra Olivia Borgmann**[2,3,4]
Jana Sommer[2,3,4], **Markus Vomhof**[2,3,4], **Cecilia Brunoni**[2,3,4]
Ismael Bruck-Ramisch[2,3,4], **Luis Enders**[2,3,4]
Andrea Icks[2,3,4], **Stefan Conrad**[1]
[1]Institute of Computer Science, Heinrich Heine University Düsseldorf
[2] Institute for Health Services Research and Health Economics, German Diabetes Center (DDZ),
Leibniz Center for Diabetes Research at the Heinrich Heine University Düsseldorf
[3]Institute for Health Services Research and Health Economics, Centre for Health and Society,
Faculty of Medicine, Heinrich Heine University Düsseldorf
[4]German Center for Diabetes Research (DZD), München-Neuherberg, Germany
julia.romberg@hhu.de

Abstract

Identifying patient information needs is an important issue for health care services and implementation of patient-centered care. A relevant number of people with diabetes mellitus experience a need for information during the course of the disease. Health-related online forums are a promising option for researching relevant information needs closely related to everyday life. In this paper, we present a novel data corpus comprising 4,664 contributions from an online diabetes forum in German language. Two annotation tasks were implemented. First, the contributions were categorised according to whether they contain a diabetes-specific information need or not, which might either be a non diabetes-specific information need or no information need at all, resulting in an agreement of 0.89 (Krippendorff's α). Moreover, the textual content of diabetes-specific information needs was segmented and labeled using a well-founded definition of health-related information needs, which achieved a promising agreement of 0.82 (Krippendorff's α_u). We further report a baseline for two sub-tasks of the information extraction system planned for the long term: contribution categorization and segment classification.

1 Motivation

Diabetes mellitus (DM) is a chronic disease with an estimated global prevalence of about 10% and steadily rising (Saeedi et al., 2019). To reduce the risk of diabetes-related acute and long-term complications, continuous health care and support as well as a high level of self-management skills are required (American Diabetes Association, 2020). A relevant number of people with DM experience a need for information during the course of the disease (Grobosch et al., 2018; Biernatzki et al., 2018). According to Ormandy (2011), a patient information need is defined as ”the recognition that their knowledge is inadequate to satisfy a goal, within the context or situation that they find themselves at a specific point in the time”. For example, shortly after diagnosis, individuals with DM report a high need for information with regard to new treatment strategies that simplify their everyday life (Grobosch et al., 2018).

To support individuals with DM adequately, identifying their information needs is an important issue in the implementation of patient-centered care (Scholl et al., 2014). Nevertheless, research in this area is still sparse. Only a few studies consider information needs during the course of the disease or in everyday life (Grobosch et al., 2018; Biernatzki et al., 2018). Studies from the field of diabetes often rely on classic quantitative or qualitative survey methods, such as questionnaires and interviews. However, to find the information they need, people with DM use online health communities, for example online diabetes forums, to improve their disease management (VanDam et al., 2017; Kuske et al., 2017; Reidy et al., 2019). An online forum can be described as an “asynchronous communication with other community members (e.g., patients can put questions to professionals or peers)” (van der Eijk et al., 2013).

We focus on the analysis of textual contributions from online diabetes forums, as personal experiences reported with classic methods may differ from personal experiences reported in online forums. An

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 19–26
Barcelona, Spain (Online), December 12, 2020.

advantage compared to traditional methods is that the communication of people with DM in online diabetes forums can highlight information needs that are more closely related to everyday life (Seale et al., 2010). This might be due to the perceived anonymity in online forums and the absence of a researcher, which can result in an openness of the forum users (Jamison et al., 2018). Nevertheless, there are several challenges for researchers using online forums as data sources. These include identifying relevant passages in a mass of posts (VanDam et al., 2017) and consideration of the more spoken language used by users when writing their posts (Seale et al., 2010). To the best of our knowledge, there is only one previous study in the field of information needs in diabetes, which uses online forums as data sources (Biernatzki et al., 2018): Ravert et al. (2004) analysed posts in online forums from adolescents with type 1 DM. They found that adolescents with type 1 DM visit online forums primarily to seek for social support, to receive information or counseling, and to exchange experiences. These findings support our motivation to explore information needs in the context of online forums.

Machine learning and natural language processing have been increasingly applied for the research of health-related issues in social media, e.g. content analysis, opinion mining, and the development of domain-specific lexicons (Denecke and Nejdl, 2009; Chen, 2012; Sudau et al., 2014; Sokolova and Bobicev, 2013; Bobicev et al., 2012; Goeuriot et al., 2012). There are two recent works focusing in particular on diabetes (Liu and Chen, 2015; Bell et al., 2018). Liu and Chen (2015) built a system for the detection and extraction of adverse drug events for drug safety surveillance. The developed approaches are evaluated on a case study corpus from a diabetes patient forum in the United States. Bell et al. (2018) investigated the automatic detection of the risk of type 2 DM directly from the Twitter activity of a person. In both cases English language data sets are used.

In this paper we present a novel corpus of German-language forum data on diabetes in which information needs are coded. The corpus was created with two objectives in mind, namely (a) the recognition of diabetes-specific information needs at contribution level and (b) the identification of relevant text segments within contributions that contain a diabetes-specific information need. Our intermediate research goal is the building of an information extraction (IE) system which consists of three steps. First, contributions with diabetes-specific information needs are filtered. Secondly, relevant text segments within the contributions with diabetes-specific information needs are identified. Third, the relevant text segments are classified according to the dimensions of patient information needs. The developed data corpus will serve as training set for the development of such a system. We also show a first approach to both classification tasks. The text segmentation step will be investigated in future work. In the long term, the IE system will allow for an automatic analysis with regard to information needs in online diabetes forums, e.g. which topics trigger increased information needs and which groups have a need, and to develop a search system that can help individuals finding relevant forum posts related to their diabetes-specific information needs.

The remainder of the paper is structured as follows. Section 2 introduces the definition of patient information needs on which we base our annotation schema. Subsequently, Section 3 outlines the data basis for the corpus, the annotation process is described in detail and the resulting corpus is analysed. Finally, Section 4 presents a basic approach for the contribution categorization and the segment classification tasks and Section 5 concludes our work and highlights the future research.

2 Ormandy's Definition of Patient Information Needs

Ormandy (2011) defines a patient information need as the "recognition that their knowledge is inadequate to satisfy a goal, within the context/situation that they find themselves at a specific point in the time". Within this definition, the following four concepts are addressed, which can either activate or influence information needs: (i) goal/purpose, (ii) context, (iii) situation and (iv) time.

(i) **Goal/Purpose:** A need for information arises from the underlying purpose of trying to achieve one or more goals, e.g. to perform self-management tasks.

(ii) **Context:** Context factors are defined according to the Wilson model of information behaviour as factors that influence the process of information seeking (Wilson, 2000). These include psychological and cognitive factors (e.g. emotions and interests), stress and coping strategies, perceived self-efficacy, demographic factors (e.g. age and gender), and role-related and environmental factors (e.g. social networks and access to health care).

(iii) **Situation:** The situation is defined as the “particular set of circumstances in which people find themselves that creates an awareness of an information need” (Ormandy, 2011). Possible situations can be events, experiences, or encounters. Examples of health-related situations are the perception of symptoms or even the perception of a life-threatening condition.

(iv) **Time:** Information needs arise at a certain point in time during an individual’s progress of disease which can be influenced by context and situation factors.

3 Data Corpus

In the following, the data source and the annotation process are presented. In addition, the reliability of the annotations is measured and the resulting data set is analysed.

3.1 Source and Preprocessing

The data corpus was constructed from the German-language online forum *forum.diabetesinfo.de*, which appears to be of reasonably high linguistic quality compared to other forums focusing on diabetes. Only publicly accessible threads were considered. When registering in the forum, forum users were made aware by the forum operator that these posts are made visible to the public. Contributions are written using pseudonyms, thus preventing direct inference to a person. Under the assumption that information needs are primarily formulated in the initial post of a thread, only the first contributions and the thread titles were included in the data set. Foreign-language posts and contributions shorter than 20 characters were excluded. HTML tags, emoticons, pictures, and applications were filtered out. Links, quotations, and tables were retained. After preprocessing, a share of 4,664 documents was used to build the annotated corpus. Additional 557 documents were used to develop the annotation guidelines.

3.2 Annotation Tasks

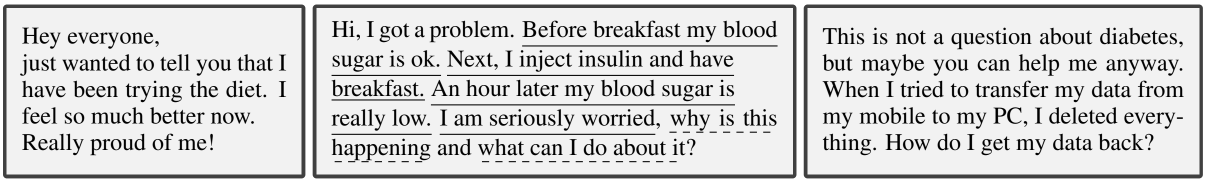
In order to create a suitable training data set for the information extraction system, two aspects must be taken into account. First, the documents of the corpus need to be coded with regard to the existence of a diabetes-specific need for information. Second, the textual content of documents in which a diabetes-specific information need is described must be segmented and the units of information must be labeled according to Ormandy’s definition.

Annotation Task 1: Document Categorization

The data set consists of documents with diverse content. Not every contribution was written to satisfy an information need, which makes it necessary to categorize the documents according to whether information needs are included or not. Contributions that fall into the latter category are, for instance, experience reports or success stories that are shared with the community. Review of the data also revealed that some information needs are not related to diabetes. Contributions with diabetes-specific information needs range from technical questions, e.g. on pumps, to nutritional issues. Typical postings for non-diabetes-specific information needs are e.g. questions about forum functions. Since our research focus is on diabetes-specific information needs, a further subdivision is made. This results in an annotation scheme on document-level with three categories: *diabetes-specific information need* (abbreviated IN_{ds}), *non-diabetes-specific information need* (abbreviated IN_{oth}) and *no information need* (abbreviated None). Figure 1 shows exemplary English-language contributions for each of the three categories.

Annotation Task 2: Detection and Assignment of the Dimensions of an Information Need

If a document contains a diabetes-specific information requirement, the IE system to be developed should extract the relevant information. In order to train this component, segments expressing the different dimensions of an information need must first be coded in the data corpus. To this end, we first discuss under which restrictions the definition of Ormandy can be transferred to health-related online forums. Analysis of the present corpus shows that the concepts goal/purpose, context and situation are represented



Hey everyone,
just wanted to tell you that I have been trying the diet. I feel so much better now. Really proud of me!

Hi, I got a problem. Before breakfast my blood sugar is ok. Next, I inject insulin and have breakfast. An hour later my blood sugar is really low. I am seriously worried, why is this happening and what can I do about it?

This is not a question about diabetes, but maybe you can help me anyway. When I tried to transfer my data from my mobile to my PC, I deleted everything. How do I get my data back?

(a) None

(b) IN$_{ds}$

(c) IN$_{oth}$

Figure 1: Exemplary contributions with document categories (a,b,c) and labeled (context/situation, goal/situation) segments of a diabetes-specific information need (b).

in the contributions. For the temporal aspect of an information need, however, this is not the case. In most cases the writer does not refer to the time her reported information need evolved. A further problem arises in the clear separation of context and situation. During the development of the guidelines it became apparent that the annotators often chose similar units for the two categories, but they often did not agree whether the dimension should be labeled as context or situation. Ormandy (2011) mentions: “A term closely related to context is situation, usually used with a narrower meaning (...)”. Since the categories cannot be clearly separated, they are hence grouped into one dimension. It is also necessary to consider how the text spans to be coded can be determined. Initially, we considered using sentences as fixed units. After reflecting on the data source, it became apparent that often different concepts are expressed in separate parts of the same sentence. As a consequence, we decided to allow the annotators to freely divide text segments on token-level within sentence boundaries. Concluding, this leads to a text segmentation task in which the resulting text segments are assigned to one of two dimensions: *goal/purpose* and *context/situation*. Figure 1b shows an annotation example in English.

3.3 Annotation Setup

Annotation guidelines were developed in sub-steps on the basis of 557 documents from the same online diabetes forum. The annotation was performed with the brat rapid annotation tool (Stenetorp et al., 2012). Following the developed guidelines, a share of 750 documents of the total 4,664 corpus documents was annotated by three coders to calculate the agreement and then adjudicated by a supervising person to build a gold standard. Due to limited capacity we decided to have the remaining 3,914 documents, hereinafter called silver standard, processed by respectively one of the three trained annotators without adjudication. However, the jointly coded part allows us to get a good estimate of the agreement between the coders. Annotations are available for download at https://dbs.cs.hhu.de/datasets/diabetesanno/.

3.4 Inter-Annotator Agreement

The reliability of annotation task 1 was measured using Fleiss’s κ (Fleiss, 1971) and Krippendorff’s α (Krippendorff,). The κ measure determines the agreement assuming that all categories are equally dissimilar. To also take into account the intuitive assumption that contributions in both categories on existing information needs are more similar to each other than to contributions without any information need, we used weighted α using the following distance function for two codings c and k:

$$d(c,k) = \begin{cases} 0, & c = k \\ 1, & c \neq k \text{ and } c,k \in \{\text{IN}_{ds}, \text{IN}_{oth}\} \\ 2, & else \end{cases}$$

Coding the documents with the three-class scheme resulted in high agreement with a Fleiss’s κ coefficient of 0.86 and a Krippendorff’s α coefficient of 0.89.

Annotation task 2 likewise scored a solid agreement. We have measured the reliability of the segmentation with Krippendorff’s α_u (Krippendorff, 1995), a coefficient for unitizing tasks. In total, an average agreement of 0.82 per document was measured. The two dimensions individually showed an average α_u of 0.82 (goal/purpose) and an average α_u of 0.79 (context/situation).

3.5 Corpus Analysis

For annotation task 1, 4,664 contributions have been coded. A single document consists of 127 tokens on average. Table 1 gives an overview of the resulting data set. In gold and silver standard, with

category	gold standard	silver standard	total
IN_{ds}	323	1,844	2,167
IN_{oth}	35	173	208
None	392	1,897	2,289
total	750	3,914	4,664

Table 1: Distribution of the contributions among the categories for annotation task 1.

dimension	gold standard	silver standard	total
context/situation	1,867	10,877	12,744
goal/purpose	672	3,679	4,351
total	2,539	14,556	17,095

Table 2: Distribution of the text segments among the dimensions for annotation task 2.

378 of 750 contributions and 2,017 of 3,914 contributions respectively, information needs are equally expressed in about half of all documents. Within these documents, the proportion of contributions containing information needs related to diabetes amounts to around 90 percent in each case. The 2,167 documents with diabetes-specific information needs taken into account in annotation task 2 consist of approximately 300,000 token, of which about 80 percent were assigned to some text segment containing either goal/purpose or context/situation. Goal/purpose mentions are fundamental for a need for information and could be identified in each document, whereas context or situation information is present in 2,044 contributions. A summary of the data set is quantified in Table 2. 17,095 text segments have been identified overall, of which about 35 percent express a goal or purpose and about 65 percent consist of a contextual or situational statement. This distribution applies to both the gold and the silver part of the corpus.

Some particularities can be found in the data corpus. The user-generated contributions show a high number of spelling mistakes, missing or unintuitive choice of punctuation and lack of semantic coherence within the texts. In addition, forum users use a very unique vocabulary, including forum-specific abbreviations for diabetes-related terms. Technical issues also seem to be central to the forum users, as technical aids (devices or software) play an important role for diabetes patients to manage and simplify their everyday life. As a result, the annotators sometimes had difficulty in understanding content, which made the annotations more challenging. In particular contributions to technical aids and medical specialities required additional knowledge to assess the extent to which these contributions are relevant to diabetes. This was also reflected in discrepancies by the annotators in the decision whether or not an information need is diabetes-specific. Some people write in a self-reflecting nature, implicitly raising information needs. Yet without an explicit statement of a goal/purpose, it is difficult to grasp the writer's intention, e.g. whether she seeks for support or not. A further difficulty in the annotation process was to decide which context and situation information is relevant to an information need.

The characteristics described above, in particular the semantic and syntactic challenges of the contributions as well as the required context knowledge and expertise for understanding them, also represent a special challenge for the IE system to be developed.

4 Baseline Approaches

In this section, we provide baselines (a) for the first step in the IE system, the contribution categorization problem, and (b) for the third step in the IE system, the classification of given text segments according to the dimensions of an information need. The intermediate segmentation step will be examined in future work. The baselines are intended to serve as a basis reference value for follow-up research. The method was the same for both tasks. Texts (either entire contributions or text segments) were split into lower case lemma tokens using *spaCy*[1] and *IWNLP* (Liebeck and Conrad, 2015) and stop words were removed. The classification was implemented with *scikit-learn* (Pedregosa et al., 2011). As classifier we chose a SVM, with fixed C of 1 and linear kernel, as SVMs are known to achieve good results for various text classification tasks. The evaluation was performed through a five-fold cross-validation. All unigrams that occurred in at least 10 documents and in at most 80 percent of the respective training data set were retained in the vocabulary, and the resulting representations were weighted with *tf-idf*.

Table 3 shows the results for both tasks. We first treat the contribution categorization task as a three-class problem. Diabetes-specific information needs are already recognised with a promising F_1 of 0.75.

[1] https://spacy.io/

	Contribution Categorization			Segment Classification	
	None	IN$_{ds}$	IN$_{oth}$	goal/purpose	context/situation
Precision	0.78	0.73	0.20	0.75	0.80
Recall	0.71	0.82	0.00	0.29	0.41
F$_1$	0.71	0.75	0.01	0.41	0.87
F$_1$ macro	0.49			0.64	
F$_1$ micro	0.70			0.76	

Table 3: Results for contribution categorization and segment classification.

The class IN$_{oth}$ is poorly hit, which is likely due to the class imbalance in the data set. While None and IN$_{ds}$ are about the same size, the class size of IN$_{oth}$ with a total of 208 instances makes up less than 5% of all contributions. Analysis of wrong predictions between IN$_{ds}$ and None shows that a frequent mistake occurs in the recognition of rather technical, often long reports describing and evaluating products (e.g. blood sugar measuring devices, pumps). These are often erroneously assigned to IN$_{ds}$ although they do not contain information needs. This might be due to the subject matter, showing similar wording to some contributions with diabetes-specific information needs. The F_1 values of 0.75 for IN$_{ds}$ and 0.71 for None nevertheless show that, despite a high degree of overlap in vocabulary, there are terms or term combinations that are class specific. Another interesting aspect relates to questions. Contrary to the expectation that explicit questions with a corresponding punctuation mark are a clear indication of the existence of an information need, this did not hold true. On the one hand, contributions with rhetorical or self-reflexive questions (primarily in narrative contributions), e.g. ”How did I come across this again?”, were recognised as IN$_{ds}$. On the other hand, quite unexpectedly, short contributions consisting mainly of goal/purpose questions were assigned None. This illustrates that the use of questions is not a specific evidence of a need for information in this data corpus. The error analysis showed that a SVM based on unigrams is insufficient for the task. Instead, a more sophisticated approach that involves the context of a contribution to decide whether words or sentences indicate information needs must be used. This difficulty becomes particularly evident in the case of questions which can take on different functions depending on the context in which they are found, e.g. expressing the goal/purpose of an information need or, conversely, giving no indication of an information need (rhetorical or self-reflexive questions).

When classifying text segments according to the dimensions of an information need, good precision values of 0.75 and 0.80 were achieved, but the recall rate was noticeably weaker. Nevertheless, the F$_1$ macro value of 0.64 shows that our model was able to learn at least some characteristics of both classes and goes beyond pure guessing. The results imply that the correct assignment of the two dimensions of an information need based on the wording of a single text segment is very difficult. The concepts of the single segments might only become clear from the overall context of all relevant segments of a contribution.

5 Conclusion and Future Work

With the long-term goal of developing an IE system for diabetes-specific information needs from forums, we introduced a German-language online diabetes forum corpus annotated on different levels. 4,664 forum posts were coded for the identification of diabetes-specific information needs on document-level. 2,167 contributions that contain a diabetes-specific information need were segmented and labeled using a well-founded definition. In this context we discussed how Ormandy’s definition of patient information needs can be applied to the domain of online forums. The resulting agreement values of 0.86 κ, 0.89 α and 0.82 α_u prove good reliability. Further, a SVM approach was applied. A promising F$_1$ of about 0.75 was achieved in the identification of diabetes-specific information needs at contribution level. For the finer classification of text segments according to the dimensions of an information need, the approach performed weaker, with an improvable F$_1$ macro of 0.64. The results offer a baseline for further work.

Following this work, we will improve the classification approaches and also work on text segmentation methods, in order to develop a complete information extraction system in the long run.

References

American Diabetes Association. 2020. Introduction: Standards of Medical Care in Diabetes-2020. *Diabetes Care*, 43(Suppl 1):1–2.

Dane Bell, Egoitz Laparra, Aditya Kousik, Terron Ishihara, Mihai Surdeanu, and Stephen Kobourov. 2018. Detecting Diabetes Risk from Social Media Activity. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 1–11. Association for Computational Linguistics.

Lisa Biernatzki, Silke Kuske, Jutta Genz, Michaela Ritschel, Astrid Stephan, Christina Bächle, Sigrid Droste, Sandra Grobosch, Nicole Ernstmann, Nadja Chernyak, and Andrea Icks. 2018. Information needs in people with diabetes mellitus: a systematic review. *Systematic Reviews*, 7:27.

Victoria Bobicev, Marina Sokolova, Yasser Jafer, and David Schramm. 2012. Learning Sentiments from Tweets with Personal Health Information. In *Proceedings of the 25th CanadianConference on Advances in Artificial Intelligence*, pages 37–48. Springer, Berlin, Heidelberg.

Annie T. Chen. 2012. Exploring online support spaces: Using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups. *Patient Education and Counseling*, 87(2):250–257.

Kerstin Denecke and Wolfgang Nejdl. 2009. How valuable is medical social media data? Content analysis of the medical web. *Information Sciences*, 179(12):1870–1880.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Lorraine Goeuriot, Jin-Cheon Na, Wai Yan Min Kyaing, Christopher Khoo, Yun-Ke Chang, Yin-Leng Theng, and Jung-Jae Kim. 2012. Sentiment Lexicons for Health-Related Opinion Mining. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 219–226. Association for Computing Machinery.

Sandra Grobosch, Silke Kuske, Ute Linnenkamp, Nicole Ernstmann, Astrid Stephan, Jutta Genz, Alexander Begun, Burkhard Haastert, Julia Szendroedi, Karsten Müssig, Volker Burkart, Michael Roden, and Andrea Icks. 2018. What information needs do people with recently diagnosed diabetes mellitus have and what are the associated factors? A cross-sectional study in Germany. *BMJ Open*, 8(10):e017895.

James Jamison, Stephen Sutton, Jonathan Mant, and Anna De Simoni. 2018. Online stroke forum as source of data for qualitative research: insights from a comparison with patients' interviews. *BMJ Open*, 8(3):e020133.

Klaus Krippendorff. Computing Krippendorff's Alpha-Reliability. *Annenberg School for Communication: Departmental Papers*. Retrieved from https://repository.upenn.edu/asc_papers/43.

Klaus Krippendorff. 1995. On the Reliability of Unitizing Continuous Data. *Sociological Methodology*, 25:47–76.

Silke Kuske, Tim Schiereck, Sandra Grobosch, Andrea Paduch, Sigrid Droste, Sarah Halbach, and Andrea Icks. 2017. Diabetes-related information-seeking behaviour: a systematic review. *Systematic Reviews*, 6:212.

Matthias Liebeck and Stefan Conrad. 2015. IWNLP: Inverse Wiktionary for Natural Language Processing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 414–418. Association for Computational Linguistics.

Xiao Liu and Hsinchun Chen. 2015. Identifying Adverse Drug Events from Patient Social Media: A Case Study for Diabetes. *IEEE Intelligent Systems*, 30(3):44–51.

Paula Ormandy. 2011. Defining information need in health–assimilating complex theories derived from information science. *Health Expectations*, 14:92–104.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Russell D. Ravert, Mary D. Hancock, and Gary M. Ingersoll. 2004. Online Forum Messages Posted by Adolescents with Type 1 Diabetes. *The Diabetes Educator*, 30(5):827–834.

Claire Reidy, David C. Klonoff, and Katharine D. Barnard-Kelly. 2019. Supporting Good Intentions With Good Evidence: How to Increase the Benefits of Diabetes Social Media. *Journal of Diabetes Science and Technology*, 13(5):974–978.

Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Belma Malanda, Suvi Karuranga, Nigel Unwin, Stephen Colagiuri, Leonor Guariguata, Ayesha A. Motala, Katherine Ogurtsova, Jonathan E. Shaw, Dominic Bright, and Rhys Williams. 2019. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Research and Clinical Practice*, 157:107843.

Isabelle Scholl, Jördis M. Zill, Martin Härter, and Jörg Dirmaier. 2014. An Integrative Model of Patient-Centeredness – A Systematic Review and Concept Analysis. *PLOS ONE*, 9(9):e107828.

Clive Seale, Jonathan Charteris-Black, Aidan MacFarlane, and Ann McPherson. 2010. Interviews and Internet Forums: A Comparison of Two Sources of Qualitative Data. *Qualitative Health Research*, 20(5):595–606.

Marina Sokolova and Victoria Bobicev. 2013. What Sentiments Can Be Found in Medical Forums? In *Proceedings of Recent Advances in Natural Language Processing*, pages 633–639. Shoumen, Bulgaria: INCOMA Ltd.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Fabian Sudau, Tim Friede, Jens Grabowski, Janka Koschack, Philip Makedonski, and Wolfgang Himmel. 2014. Sources of Information and Behavioral Patterns in Online Health Forums: Observational Study. *Journal of Medical Internet Research*, 16(1):e10.

Martijn van der Eijk, Marjan J. Faber, Johanna W. M. Aarts, Jan A. M. Kremer, Marten Munneke, and Bastiaan R. Bloem. 2013. Using Online Health Communities to Deliver Patient-Centered Care to People With Chronic Conditions. *Journal of Medical Internet Research*, 15(6):e115.

Courtland VanDam, Shaheen Kanthawala, Wanda Pratt, Joyce Chai, and Jina Huh. 2017. Detecting clinically related content in online patient posts. *Journal of Biomedical Informatics*, 75:96–106.

Thomas D. Wilson. 2000. Human Information Behavior. *Informing Science*, 3(2):49–56.