GDWDS: First Insights from a Student-based Key Phrase Annotation Process of Medical Information Needs on a Novel German Diabetes Web Data Set

Julia Romberg
Institute of Computer Science
Heinrich Heine University Düsseldorf
D-40225 Düsseldorf, Germany
romberg@cs.uni-duesseldorf.de

ABSTRACT

The information needs of individuals are at the forefront of various issues. One platform that users use to address their needs is Internet forums. Medical forums in particular are very much shaped by questions and articulated needs.

As part of our research, the need for information is to be examined specifically in the context of diabetes expressed in web forums. For this purpose we introduce GDWDS, a novel German diabetes web data set. Assuming that the information needs can be understood as key phrases, the record was annotated by student annotators. Three tasks were addressed: First the recognition of key phrases in a document. Second, the annotators were requested to summarize key phrase of the same content in one group. Third, every group should be represented by the most meaningful key phrase contained in this group.

The main annotation task of identifying the text units that express information needs lead to an average Krippendorff's unitized Alpha of 0.439 which is promising. The tasks of grouping the key phrases and selecting a representative could only be evaluated to a limited extent due to their subjective dependence on the key phrase detection task.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications: data mining; I.2.7 [**Artificial Intelligence**]: Natural Language Processing: language parsing and understanding, text analysis; H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Information Retrieval, Information Needs, Keyphrase Extraction

30^{th} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 22.05.2018 - 25.05.2018, Wuppertal, Germany.

1. INTRODUCTION AND MOTIVATION

Nowadays social media has taken an important place in most people's lives. Platforms such as Twitter, Facebook and Instagram are widely used to communicate feelings and opinions. Besides the just mentioned prominent examples exists a variety of other mediums which serve as speaking tube. Especially blogs and forums are used to inform about specific themes and to discuss them.

One particular aspect that is increasingly picked out as a central theme are health-related topics. In [16] Sokolova et al. have identified multiple reasons for the use of medical forums in several studies from the years 1990 to 2009: On the one hand, persons who are either suffering theirselves from a disease or whose beloved ones do, may search for information that exceeds the information provided by an attending doctor. Thereby, the information need ranges from psychological, physical, and social aspects of treatments to alternative treatments. On the other hand, forums offer a point of contact for people that seek for emotional support, especially from other fellow sufferers. Furthermore, forums often provide a feeling of anonymity to members, which helps them to communicate more openly about their experiences.

A widespread disease is the metabolic disease diabetes mellitus. In 2017, according to the International Diabetes Federation[1], approximately 425 million adults worldwide have suffered from Diabetes[2], which is more than 5% of the world population.

Diabetes appears mainly in two different forms, Type 1 and Type 2. While genetics and environmental factors are mostly held responsible for Type 1, Type 2 is additionally associated with lifestyle factors. Diabetes is a disease which often accompanies the affected persons their entire lives. In order to facilitate that these persons can live a normal life nevertheless, a good insulin adjustment and an appropriate routine in exercise and nutrition may be needed. Institutions, for example the *Deutsches Diabetes-Zentrum*[3] (German Diabetes-Center), aim to improve patients' quality of life, among other things, by focusing on the patient's information needs and preferences. Patient statistics on these points are collected using questionnaires. This course of action unfortunately shows some weaknesses: (i) The number of questions is limited. (ii) Only a limited number of people can take part in a survey. (iii) The evaluation is time-consuming and diffi-

[1] http://www.idf.org/
[2] www.diabetesatlas.org
[3] http://ddz.uni-duesseldorf.de/en/

cult, especially when having free-text fields, which currently require a (manual) qualitative analysis. (iv) The physicians and researchers developing the questionnaires usually have another point of view on the diseases and how the treatments affect the patients. Finally, the researchers who develop the questionnaires usually have a different perspective on a disease and on how a treatment affects the patients than those affected. Therefore, patient-relevant questions could be omitted.

An alternative approach for the analysis of information needs and preferences of diabetes patients is the use of information retrieval techniques. A first intuition would be to apply natural language processing techniques to the questionaries' included free-text fields. However, to address all the problems listed above, the whole course of action should be changed: Instead of manually posing questions and manually analyzing them in tedious work, existing resources can be used, namely medical online forums. At this point, it is necessary to discuss if and to what extent an online forum community can represent the general population of diabetic affects. In [7] online social networking for diabetes is examined. The authors found in the study that online groups on diabetes, using the example of Facebook, cover a broad spectrum of involved persons, such as patients and their families. Another interesting fact is a special technical affinity of diabetes patients, which is due to the current treatment methods such as app-based monitoring of diabetes. This suggests that the existing information needs of the total population are reflected to a large extent in online health media. At the same time, however, it is important to remember that older patients or patients who have been in treatment for a very long time are unlikely using these channels. It must also be taken into account that the data corpus of this work refers to the information needs in industrialized countries using the example of Germany.

In this paper, we focus on the annotation process of a data corpus based on forums of this kind. Our long-term research goal is the automated recognition and extraction of information needs. In order to be able to implement this task well-founded, a prior focus on an appropriate corpus annotation is necessary as evaluation is an essential point to keep in mind. The remainder of this paper is structured as follows: First, the data set is introduced. The implementation and nature of the annotation and the different steps of the annotation process are explained. Subsequently, the quality of the resulting data set is calculated and discussed by means of an Inter-Annotator Agreement. We then conclude and describe the use of this study for a further annotation process.

2. RELATED WORK

There has been previous research in the field of social media health and diabetes in the recent years.

Multiple publications have focused on content analysis on medical social media texts. Denecke et al. [4] compared different social media health data sources by first extracting medical concepts and then pointing out content differences. They also focused on the binary classification problem of informative versus affective statements. In [3] medical support group texts were clustered into topics whereas in [17] clustering was used to analyze user preferences for the use of information sources and to analyze the users' general posting behavior. Ravert et al. [14] analyzed the content online forum messages from adolescents with Type 1 diabetes. In doing this, a corpus consisting of 340 posts was annotated with respect to age, gender, date and duration of illness. They found that diabetes affected persons visit online forums mainly for the sake of social support, information and advice along with shared experiences. These findings support our motivation of investigating the information needs in diabetes online forums. Although the corpus is interesting, the annotations unfortunately lack reference to information needs and key phrases.

A lot of research has been conducted in sentiment analysis and opinion mining [16, 15, 2, 6, 1]. The used corpora vary from medical forum data for *In Vitro Fertilization* and *Hearing Loss* over drug reviews to Twitter messages and message boards. Reader-based as well as author-centric annotation models were applied. Furthermore, domain specific lexicons were developed: In [6] a lexicon was built from drug reviews. Sokolova et al. [16] introduced *HealthAffect*, a domain-specific affective lexicon. Both papers conclude that general sentiment and affective lexicons cannot adequately serve for social media health texts because of the specific terms and language used in this area.

Further research was conducted on the detection and extraction of adverse drug events in social media texts. Karimi et al. [8] developed *CADEC*, a annotated corpus of adverse drug events. Liu et al. [11, 10] investigated on identifying adverse drug events and implemented an information extraction system for adverse drug events, both on a data set focused on diabetes. These corpora contain information specific to adverse drug events, which at best expresses a subset of the general need for information.

3. THE CORPUS

In this section the creation of an appropriate data corpus, needed for later research, is discussed. To the best of our knowledge, there is no existing data corpus consisting of diabetes forum messages that has been annotated in a sense we could use for our analysis.

Our objective is the recognition and extraction of the information needs of forum users. The following pattern was recognized in forum posts. A contribution is opened up in the multiplicity of cases in order to ask of the community information on a certain topic or an answer to a concrete question. For this purpose, first the more detailed circumstances are explained and then the corresponding questions are formulated. Subsequently, other users respond to the post with answers and descriptions of their own experiences.

The data corpus GDWDS was build on a freely accessible German diabetes forum. The data set for this initial study was build by extracting 150 forum contributions from the corpus. Assuming that a user announces his information needs when creating a thread, only the initial contributions were retained while the replies were discarded. Often the title of a thread also contains important information. In order to keep this information, the thread title was added to the document as a heading.

3.1 Annotation Setup

We see the problem of recognizing information needs as an information extraction problem. Key words and key phrases expressing these needs should be extracted to allow a summary of the information needs. To form a gold standard for the evaluation of such techniques, an annotation of the da-

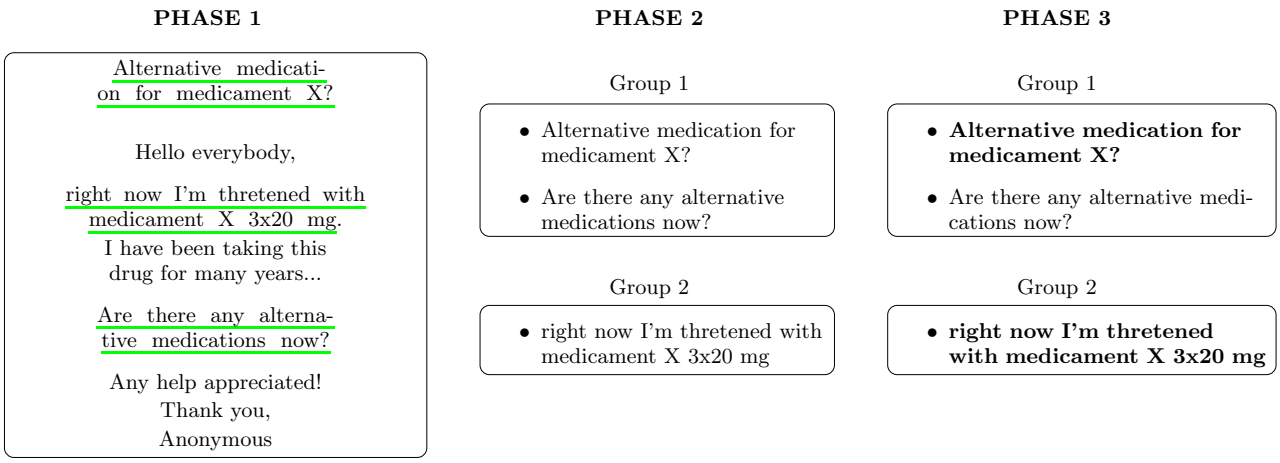

Figure 1: The three annotation phases (phase 1 - key phrase recognition, phase 2 - key phrase grouping, phase 3 - best key phrase identification) are illustrated by means of an example document.

ta set must be made. A text sequence is to be divided into annotation units, which are then assigned to a class. According to our task there are two classes: *key phrase* and *no key phrase*.

Twenty-five student annotators were divided into five annotation groups with 4 persons and one group of 5 persons. The GDWDS's 150 documents were divided among the six groups so that each group had to handle a workload of 25 documents. The annotators were instructed to carry out the annotations independently. The annotations were implemented using MDSWriter [12]. This tool, originally developed for creating multi-document summarization corpora, was used in a modified form. We only use the first three phases of the tool: recognizing key phrases, grouping key phrases with the same content, and identifying a best key phrase within each group.

In an introductory phase the annotators were explained the guidelines to be fulfilled. These guidelines were developed based on the guidelines of [12].

It should be noted that the annotation process presented here is rather unusual. In most cases, in a qualitative annotation setting with extensive rules, only a subset of the data is processed by several annotators in order to be able to estimate the annotation quality. The remaining corpus is divided on the individual annotators. In the study presented here, the focus is on testing the admissibility and completeness of the guidelines that have already been developed. The results contribute to the final annotation of the entire body.

3.1.1 Phase 1 - Key Phrase Recognition

1. The participants were first requested to read a document, i.e. a forum contribution, completely before starting the annotation. Unknown words should be looked up or asked in advance to ensure comprehension.

2. Subsequently, the key phrases should be marked. The following guidelines should be followed:

   - A key phrase should at least consist of a predicate plus subject or a predicate plus an object.
   - A key phrase must not exceed a sentence boundary.
   - A key phrase is intended to contain important content related to the information need expressed in the document. This may refer to an explicit formulation as a question but also to contextual information that is important for an accurate description of the information need.
   - If a key idea is described several times in the document, all entries must be marked.

3. Finally, the recognized key phrases should be reviewed and checked for clarity, accuracy and content.

The clarification of unknown words in (1.) is of particular importance, since the medical context requires many technical terms and abbreviations. In addition, there are a few abbreviations that differ from the conventional vocabulary. These terms seem to have evolved within the forum community.

3.1.2 Phase 2 - Key Phrase Grouping

Following the identification of the key phrases, the participants were asked to group phrases of the same content together. Although the texts to be annotated are on average only 1187 characters long, re-mentions occur, among other things caused by the addition of the thread title.

3.1.3 Phase 3 - Best Key Phrase Identification

In the final annotation phase, participants should select a representative in each group of key ideas. The representative should contain the largest possible information content.

The three annotation phases are illustrated in Figure 1. In phase 1, the annotator sees the document to be unitized. The selected key phrases are underlined in green. Subsequently the key phrases are requested to be grouped according to their content. In this example two key phrases refer to the need for information in relation to an alternative medication to the current one. Hence, they are summed up. The third key phrase relates to content information, which clarifies the expressed information needs and is equally important. This

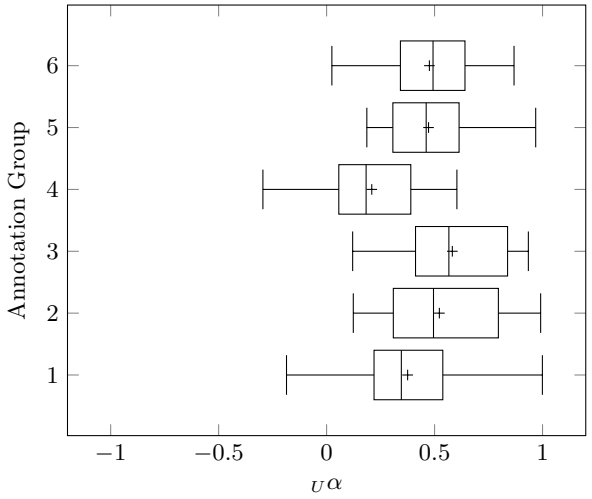

Figure 2: Box plots depicting the distribution of the achieved values ($_U\alpha$) within the individual annotation groups.

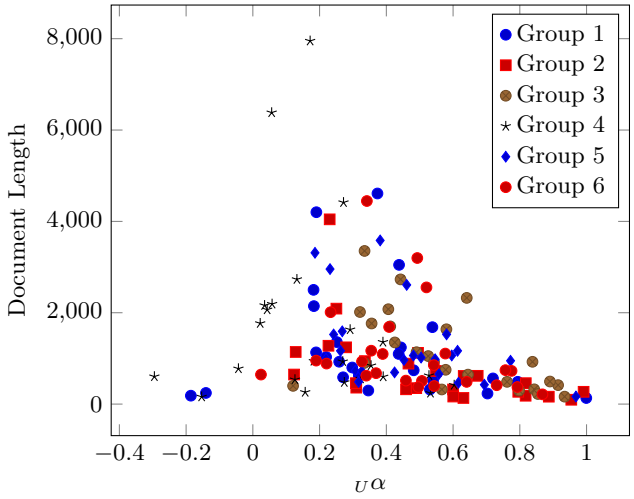

Figure 3: Plot showing the relation of U_α and the document length.

phrase builds a second group of key phrases. Finally in phase 3 the annotator must decide for a best key phrase inside of each group created in the previous phase. The best key phrase is bold. For group 2 no discussion is needed. The representative in group 1 is selected based on the request for the largest possible information content as not only the question for an alternative but also the name of the currently used medicament is stated.

3.2 Inter-Annotator Agreement

Following the annotation task itself, the resulting annotations need to be evaluated. For this the Inter-Annotator Agreement of the persons of the same annotation group is calculated.

3.2.1 Phase 1 - Key Phrase Recognition

Since annotation phase 1 is a unitizing task with one category, we use Krippendorff's unitized alpha $_U\alpha$ (introduced in [9]) as a measure. $_U\alpha \in [-1, 1]$ describes the correspondence of different annotators' coding units on the same text document. 1 expresses maximum agreement, 0 shows that no correlation exists between the units and the classes, and -1 symbolizes a uniform disagreement. The calculations were carried out with DKPro Agreement [13].

First, for every of the six groups of annotators described in Section 3.1 the groups' agreement over all 25 documents was considered. Table 1 shows the agreement within the annotation groups. Annotation group 2 and 3 obtain the best agreement having an $_U\alpha$ above 0.5. Group 1, 5 and 6 agree with a value greater than 0.4. Group 4, however, performs significantly worse achieving only an $_U\alpha$ of 0.210. One possible explanation might be the text length of the documents. The average length of a text document in group 4 was 1645.36 characters. The other groups had on average shorter texts with at least 400 characters less. The longer a text is, the more descriptive the information requirement is described. Likewise, increasingly diverting content may occur. This makes unitizing key phrases harder. Nevertheless, the remaining groups achieve encouraging Inter-Annotator Agreements.

To evaluate the annotations more accurately and in more detail, the Inter-Annotator Agreements are further examined at the document level. The quality of the annotation results of the individual documents is illustrated in Figure 2. The box plot of each group shows the worst as well as the best $_U\alpha$ value achieved for a document assigned to this group. The boxes illustrate the quartiles and the median. The mean value shown in Table 1 is illustrated with a cross. As can be seen, the agreement within the groups is very variable. The box plots reflect again that group 4 performs worse than the other groups. However, the values achieved in every group extend over an interval of length 0.8 to 1.2, which corresponds to approximately half the value range of Krippendorff's unitized Alpha. Although the average mean value of agreement of 0.439 appears acceptable across the six groups, the large variability of the data indicates that annotation quality must be considered with caution.

Figure 3 illustrates the relation between the document length and the agreement. Unexpectedly, the assumption that longer documents lead to a worse agreement is not confirmed here. Although a slight tendency is visible, both the left and right tail of the distribution represent short documents. Accordingly, the content of the marginal documents was analyzed. In documents with poor agreement, it was noted that the annotators were often in agreement on important information. However, the distribution of this information into the different key phrases was solved very differently. Especially in terms of conjunctions like „and, or, ... “ the annotators were divided. Some annotators split into more granular units than the others. The importance of context information was also assessed differently. For example, in one document a patient described a need for information against the background of his type 1 illness. He also stated since when he was affected by the disease. Here, the annotators were divided over whether the temporal context is important for the formulation of the information need. At this juncture, it should be remembered that the students had no domain knowledge in the field of medicine or diabetes, making it difficult to make a reasoned decision. Furthermore, annotation errors were observed. In some annotations, the

Annotation Group	1	2	3	4	5	6
${}_U\alpha$	0.409	0.505	0.523	0.210	0.443	0.429

Table 1: Showing the Inter-Annotator Agreement by group according to ${}_U\alpha$.

content doubling of a key phrase has not been re-marked. Individual annotators tended to classify the key phrases so finely that the selected key phrases individually could not express a key content of the text.

3.2.2 Phrase 2 - Key Phrase Grouping and Phase 3 - Best Key Phrase Identification

Following the recognition of key phrases, the subsequent grouping needs to be revised. Due to the dependency on phase 1, it is difficult to assess whether the same groupings have been made. In an optimal scenario, starting from an equal set of key phrases, an Inter-Annotator Agreement could be calculated for the same coding units and a set of classes corresponding to the number of key phrase groups.

In order to be at least partially able to analyze how much the annotators in phase 2 agree in their decisions, only documents with an alpha greater than 0.439 (corresponding to the average mean of phase 1) are considered. Furthermore, documents whose annotators disagree on the number of units are excluded from consideration. With these restrictions, we want to make sure that the same phrases were detected in phase 1 allowing a small variation tolerance in order to build a suitable initial data situation for phase 2. Thus, we can measure the Inter-Annotator Agreement for these documents. As appropriate measures we use simple percentage agreement PA on the one hand and Fleiss Kappa κ [5] on the other hand. We use DKPro Agreement for the calculation again. Unfortunately, only three documents meet the required criteria. For the first of them, all annotators only assigned one key phrase unit. Accordingly, there is only one group of key phrases and thus, both PA and κ are 1. The second document that fulfills the criteria contains according to phase 1 two key phrase units. Every annotator summarized them into the same group which leads to a perfect agreement in terms of both measures. Prevailing phase 3, it is to be noted that also the best nuggets were equally chosen. The last document consists of three units. While three of four annotators completely agreed in dividing the units into two groups and making the same assignments, the fourth annotator fixed on three groups which finally lead to a PA of 0.66 and $\kappa = 0.38$. The three annotators building the same groupings did, however, not agree on the best key phrase per group.

Since it is obvious that the quality of the dependent annotations can only be analyzed to a limited extent and therefore not very meaningful, we will not go further into phase 2 and 3 here.

4. CONCLUSION AND FURTHER WORK

In this work we presented an annotation study of medical information needs on a german diabetes data set. Student annotators were instructed to detect key phrases, group them according to similar content and then to find a representative key phrase for every group. For this, the students had to follow guidelines, presented in Section 3.1.

Subsequently, the obtained annotations were evaluated. Since there obviously is no gold standard, in the evaluation part we focused on the Inter-Annotator Agreement. The results for phase 1 are promising. Although there is an obvious variance in the data, for almost half of the documents the annotators agreed with at least 0.5, annotation group 4 excluded. We observed different types of problems. The lack of subject-specific knowledge was one of the main problems annotators had to face. A second problem was the different view of a key phrase's granularity level. Finally we detected some cases in which the annotators did not concentrate on the given guidelines producing poor annotations. Phase 2 and 3 could not be investigated meaningfully as phase 1 directly conditions the initial data situation of the other phases.

These findings lead us to the assumption that, in order to further increase the quality of the data corpus, experts need to be taken into account. Healthcare professionals are important but likewise social media experts are of interest because of the particular vocabulary that is used in forums. Our data set showed very specific expressions that are only used in the online context of diabetes. Another fact we must address are the guidelines. These need to be revised with regard to the annotators insecurities concerning key phrase granularity, the rules for key phrase grouping and for choosing a representative key phrase. The annotators also reported that the very subjective nature of the texts was another difficulty.

Summarized this first annotation approach on the GDWDS achieved promising results, especially in the main phase, phase 1. As the students had only a short introductory class into the annotation task, this approach can be seen as a crowd-sourcing attempt. However, to further increase annotation quality, we see an expert-based approach at an advantage. In future work this issue will be addressed. Phase 2 and 3 should be neglected until phase 1 produces results with a quality sufficient for the further phases. Nonetheless the development of appropriate measures for dependent annotation tasks may be an interesting area of research.

If the annotation quality of the GDWDS is ensured, the actual process of keyphrase extraction can be started. As a first step we plan to apply and evaluate state-of-the-art algorithms for key phrase extraction on GDWDS. Machine learning algorithms and deep learning approaches are prevalent in this field. For example, in [18] an interesting approach to keyword extraction from Twitter using recurrent neural networks is presented. Alternatively, rule-based or graph-based approaches should be considered. In accordance with the results it can be evaluated whether these existing techniques can be applied to our problem. Depending on the results, new algorithms may then be developed.

5. ACKNOWLEDGMENTS

We want to thank the student annotators, namely Deniz Ates, Bashkim Berzati, Christian Born, Markus Brenneis, Nurhan Chahrour, Björn Ebbinghaus, Julia Fischer, Andreas Funke, Philipp Grawe, Frederik Grieshaber, Tobias Alexander Hogrebe, Michael Janschek, Moritz Kanzler, Sergej Korlakov, Daniel Laps, Johannes Müller, Alexander Ober-

straß, Karsten Packeiser, Kevin Robert Pochwyt, Regina Stodden, Emil Warkentin, Dennis Weber, Susanna Welzel, Julian Zenz and Milos Lukas Ziolkowski.

6. REFERENCES

[1] T. Ali, D. Schramm, M. Sokolova, and D. Inkpen. Can i hear you? sentiment analysis on medical forums. In *IJCNLP*, pages 667–673, 2013.

[2] V. Bobicev, M. Sokolova, Y. Jafer, and D. Schramm. Learning sentiments from tweets with personal health information. In *Canadian Conference on Artificial Intelligence*, pages 37–48. Springer, 2012.

[3] A. T. Chen. Exploring online support spaces: using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups. *Patient education and counseling*, 87(2):250–257, 2012.

[4] K. Denecke and W. Nejdl. How valuable is medical social media data? content analysis of the medical web. *Information Sciences*, 179(12):1870–1880, 2009.

[5] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[6] L. Goeuriot, J.-C. Na, W. Y. Min Kyaing, C. Khoo, Y.-K. Chang, Y.-L. Theng, and J.-J. Kim. Sentiment lexicons for health-related opinion mining. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 219–226. ACM, 2012.

[7] J. A. Greene, N. K. Choudhry, E. Kilabuk, and W. H. Shrank. Online social networking by patients with diabetes: A qualitative evaluation of communication with facebook. *Journal of general internal medicine*, 26(3):287–292, 2011.

[8] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81, 2015.

[9] K. Krippendorff. On the reliability of unitizing continuous data. *Sociological Methodology*, pages 47–76, 1995.

[10] X. Liu and H. Chen. Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums. In *International Conference on Smart Health*, pages 134–150. Springer, 2013.

[11] X. Liu and H. Chen. Identifying adverse drug events from patient social media: A case study for diabetes. *IEEE Intelligent Systems*, 30(3):44–51, 2015.

[12] C. M. Meyer, D. Benikova, M. Mieskes, and I. Gurevych. Mdswriter: Annotation tool for creating high-quality multi-document summarization corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 97–102, Berlin, Germany, August 2016. Association for Computational Linguistics.

[13] C. M. Meyer, M. Miesked, C. Stab, and I. Gurevych. Dkpro agreement: An open-source java library for measuring inter-rater agreement. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations (COLING)*, pages 105–109, Dublin, Ireland, August 2014. Association for Computational Linguistics.

[14] R. D. Ravert, M. D. Hancock, and G. M. Ingersoll. Online forum messages posted by adolescents with type 1 diabetes. *The Diabetes Educator*, 30(5):827–834, 2004.

[15] M. Sokolova and V. Bobicev. Sentiments and opinions in health-related web messages. In *RANLP*, pages 132–139, 2011.

[16] M. Sokolova and V. Bobicev. What sentiments can be found in medical forums? In *RANLP*, volume 2013, pages 633–639, 2013.

[17] F. Sudau, T. Friede, J. Grabowski, J. Koschack, P. Makedonski, and W. Himmel. Sources of information and behavioral patterns in online health forums: observational study. *Journal of medical Internet research*, 16(1):e10, 2014.

[18] Q. Zhang, Y. Wang, Y. Gong, and X. Huang. Keyphrase extraction using deep recurrent neural networks on twitter. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 836–845, Austin, Texas, November 2016. Association for Computational Linguistics.