

RESEARCH

Open Access



# Computer aided sentiment analysis of anorexia nervosa patients' vocabulary

Dominik Spinczyk<sup>1\*</sup> , Karolina Nabrdalik<sup>1</sup> and Katarzyna Rojewska<sup>2</sup>

\*Correspondence:

dspinczyk@polsl.pl

<sup>1</sup> Faculty of Biomedical Engineering, Silesian University of Technology, Roosevelta 40, Zabrze, Poland  
Full list of author information is available at the end of the article

## Abstract

**Background:** Diagnosing and treating anorexia nervosa is an important challenge for modern psychiatry. Taking into account a connection between the mental state of a person and the characteristics of their language, this paper presents developed and tested method for analyzing the written statements of patients with anorexia nervosa and healthy individuals, including the identification of keywords.

**Methods:** Due to the short nature of the texts, which is related to the difficulty of expressing oneself about one's body when suffering from anorexia, the bag of words approach was used for documents' information representation. The document is represented as a vector, where its various elements indicate the number of individual words. Then, a rule-based model was created, where as a collection of rules, dictionary files were used corresponding to three groups of positive, negative and neutral sounds for each subcategory. Next in the analyzed texts were searched and counted keywords. Based on the keywords found, each of the documents was categorized into one of the groups in every subcategory.

**Results:** It is possible to indicate a set of characteristics sentiment for every person. Additionally, the results of specific patient could be analyzed in six specific subcategories: self-esteem, acceptance of the assessment of the environment, emotions, autoimmune, functioning of the body and body image.

**Conclusions:** The described analysis indicates the existence of a relationship between the mental state of the author's textual health and the vocabulary he or she uses. It is possible to indicate a set of characteristic sentiment terms specific to a given group of people. Their presence is related to the author's mental state and their body image. It could help focus on specific topics during therapy.

**Keywords:** Anorexia nervosa, Sentiment analysis, Sentiment dictionary

## Background

Anorexia (AN anorexia nervosa) is an illness without a fully understood etiology characterized by a desire to lose weight through excessive restriction of food supply and disturbing self-perception body. Patients are constantly focused on their diet to control their weight. Regardless of health and even after being significantly underweight, they do not give up a restrictive diet, which can cause severe exhaustion and even death [1]. The disease may be a one-time or chronic episode, with periods of remission and recurrence.

People with eating disorders give their body, especially its mass and shape, too much importance, which manifests itself in excessive concentration on the body's silhouette

and food, often taking the form of a surplus idea [2]. There are many factors involved in shaping this particular attitude towards one's body including patient experience and personality traits. According to numerous studies, the patient's body image while suffering from AN is disturbed: they perceive their body as more powerful than it is in reality, which in affects the imposition of dietary restrictions [3–5]. The disturbed body image in AN patients may have a negative impact on the quality of food intake during the symptomatic recovery period and after restoration of normal body weight. It has been observed that this promotes the maintenance of amenorrhea [6, 7].

The research by Rojek and Opoczyńska [8] reveals the importance of developing a different body image for anorexia-affected women and what cognitive, behavioral and emotional factors can sustain and enhance their dietary behavior. Based on a comprehensive analysis of interviews with AN patients, the researchers identified the following points: fixation, aesthetic experience, object of control, object of aggression, object of devaluation, area of expression of femininity, expression of objection, method of attention, communication tool. In this context, food preferences of patients related to their perception of their body appear to be a way of realizing specific emotional and cognitive needs and their deprivation in patients with AN. Hence, it seems that the methods of evaluating the judgments and self-assessments of the body and emotions associated with it in patients with anorexia nervosa is important.

There is a connection between the mental state of a person and the characteristics of the language they use in their statements. Studies show significant differences between the theories created by healthy people and those made by people with mental disorders. The main differences in the two groups are the content of positive words and negative emotions and the use of the pronoun "and" [9]. These features are probably due to the dominant themes in the author's minds and the positive or negative interpretation of ambiguous events.

Among research works focusing on the characteristics of language, identifying features of people with different mental states can be found [10] when comparing the language used in texts about self-presentation. Persons identifying as pro-anorectic and anorexics patients. The approach proposed by the authors was based on the symbolic-statistical method of analysis used words on three main levels: grammar, psychological and thematic. Taking into account the before-mentioned features, differences between people from each group were identified. The theses of pro-anthropic tendencies were more positive than those of the second group. Texts from anorectic patients were characterized by greater features that point to anxiety and self-care.

The analysis and interpretation of speeches from anorexia patients are described in [11]. The authors classify the categories according to the aspects of corporeality and characterizing the main aspects of perceiving one's own body. These categories have been presented as follows: body as a goal (fixation point, object of aesthetic experiences, control object, object of acceptance, object of aggression) and body as a way to achieve the goals (the word of objection, a way of paying attention, a way of gaining recognition, defense against others, a communication tool).

A literary analysis indicates that language analysis may reveal certain characteristic features in statements from healthy people and those suffering from mental disorders such as AN. The article was proposed to develop and test a method for analyzing the

written statements of patients with anorexia nervosa and healthy individuals, including the identification of keywords.

## **Methods**

Natural language can be subdivided into tangible and emotional language [12]. Tangible language describes the facts or ideas in an unambiguous and relatively objective way. Emotional language, on the other hand, is based on expressing feelings, and its tone may depend on the author and is often ambiguous. In a scientific approach, the tangible type seems more appropriate for analysis, because of the simplicity of the meaning of words. Emotional type expresses sentiment—the feelings of the author. Its analysis involves additional difficulties and requires more complex algorithms or methods for their unification. However, it allows for additional information which would be impossible to obtain using only simple, factual language.

The main goal of the study is to develop and test a method for analyzing written statements from patients with anorexia and healthy people including the identification of keywords, and includes:

- Information extraction from the analyzed texts,
- Choosing the optimum form of data representation,
- Sentiment analysis,
- Comparison of results for processed and unprocessed documents and by AN patients and healthy patients.

### **Information extraction from the analyzed texts**

After the literature analysis, presented in "Background" section, and accounting for treatment of AN, six sub-categories were defined:

- Self-esteem: aesthetic way of perceiving your body,
- Acceptance of the assessment of the environment: reception of the person perceiving it by the environment,
- Emotions: experienced emotions,
- Autoimmune: descriptions of aggressive and self-aggressive behaviors,
- Functioning of the body: description of the functioning of the body,
- Body image: image of individual parts of the body.

For each subcategory, positive and negative dictionaries were used. To do this, experts assigned words from the general vocabulary of sentiment, created by Wilson, Wiebe and Hoffman [13], to the particular categories. It contains more than 8000 terms with specific polarity (positive, negative or neutral). These are nouns, verbs, adjectives and adverbs. It was created and was developed by a group of experts who evaluated the objective sound of the words in it. The dictionary is composed of English terms, so it was necessary to transform it so it could be used in the analysis of texts in the Polish language.

**Choosing the optimum form of data representation**

In order for text data to be analyzed, it is essential to select the correct text representation. One of the approaches is data representation in vector space: the text is converted to numerical vectors based on the number of the most important lexical elements present in it [14]. Examples of this approach to the problem:

- Representation based on words in the document (bag of words), where each component represents the word. This approach excludes analysis of grammar and the distance between words, destructing text to be simpler for analysis by computer,
- Representation based on the choice of sentences (bag of phrases), understood as sequences of words. The sentences that are more likely to bring meaning are important.

Due to the short nature of the texts, which is related to the difficulty of expressing oneself about one's body when suffering from AN, if the bag of words approach was used.

**Information about the frequency of terms**

The document is represented as a vector, where its various elements indicate the number of individual words. The transformation of the body of the input documents to this form requires the following:

- At the beginning of the process to remove punctuation and create a list of independent occurrences of words for each document,
- Remove words irrelevant from the point of view of further analysis (the words they form the so-called stop-list),
- Transformation of words to their basic form (called stemming). Depending on the specific language transforming words into their basic forms can be implemented using rules or dictionaries.

Words that are left after initial processing of documents are called terms. A collection of terms with all of the documents is called a dictionary. Frequency list of occurrences are then combined into a single list forming a frequency matrix  $A$ :

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & a_{ij} & \dots \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$$

where:  $a_{ij}$  element of the frequency matrix,  $m$  the number of rows of the frequency matrix corresponds to the number of terms included in the matrix frequency,  $n$  the number of columns in the array of frequency, corresponds to the number of documents of corpus.

The column number represents the index of the input document and row number represents term's index. Value of each matrix element corresponds to the number of occurrences of specific term in the document. The advantage of this approach is the simplicity of implementation of the calculation and a large range of processing methods. A

significant drawback of frequency matrix is limitation of information about each term to the number of its occurrences.

### **Sentiment analysis**

Acquisition of textual information is based on text mining techniques, methods of processing, and representation of textual data. It is based on the definition of text polarization (positive or negative), which is the simplest approach to the problem. In addition, the degree of severity may be determined to consider a wider range of emotions.

The problem of sentiment analysis can be taken in different ways. Generally, emotional intelligence algorithms use a categorical emotion model (assuming the existence of several basic, discrete, and innate emotions) and focus on classifying the text into one category. Sentiment analysis is therefore a method of categorizing text [15]. A characteristic feature of this type of document classification is the relatively small number of groups, compared to the traditional categorization approach: they are often two or three groups (“positive”, “negative” and sometimes “neutral”). As mentioned earlier, for each subcategory, positive and negative dictionaries were used. In the presented approach, a method based on lemmatization and stemming was used, which uses algorithms to bring the word to its basic form (both in terms of variation and word formation). With this approach, any use of a word, regardless of the form or variety used, is considered to be the same term. This avoids redundancy of dimensions. The sentiment of the document  $S$  is calculated as:

$$S = \frac{neg - pos}{neg + pos + neu} \quad (1)$$

where:  $neg$  the number of negative keywords in the document,  $pos$  the number of positive keywords in the document,  $neu$  the number of neutral keywords in the document.

In the first step (analysis of documents using a dictionary), a rule-based model was created. As a collection of rules, dictionary files were used corresponding to three groups of positive, negative, and neutral sounds. Next in the analyzed texts were searched and counted keywords. Based on the keywords found, each of the documents was categorized into one of the groups in every subcategory. If the number of positive terms is greater, equal or less than the number of negative terms, the document is positive, neutral or negative. Where no keyword was found, the sentiment of the document is indefinite.

### **Input data set**

Texts on which the analysis described in this paper was performed, there are the speeches of 67 people on “My body”. These statements were made in written form and divided into two groups:

- *Research group* Statements of ill persons (15 documents—one for every person). The pilot study included patients hospitalized in the Department of Pediatric Endocrinology, Hospital No. 1 in Zabrze. The study involved 15 girls aged 12–17 years with diagnosed type restriction anorexia who were hospitalized due to exacerbation of the disease and significant weight loss (BMI SDS below  $-1$ ). All patients and their

caregivers expressed a written consent for the study, which was performed anonymously.

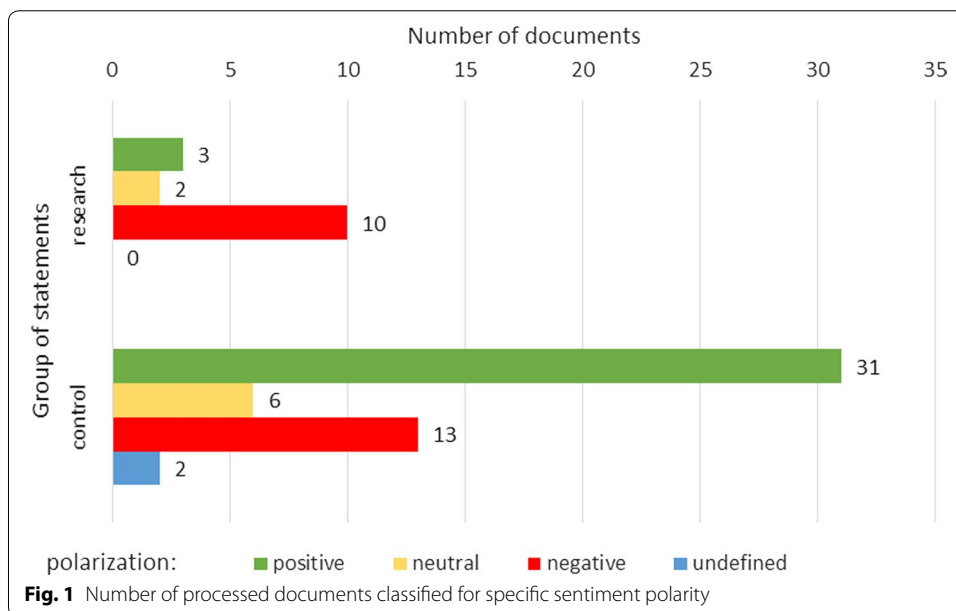
- *Control group* Statements of healthy people (52 documents). The control group consisted of healthy girls aged 13–17 years with normal body weight. The research was carried out anonymously at the Secondary School No 5 in Gliwice.

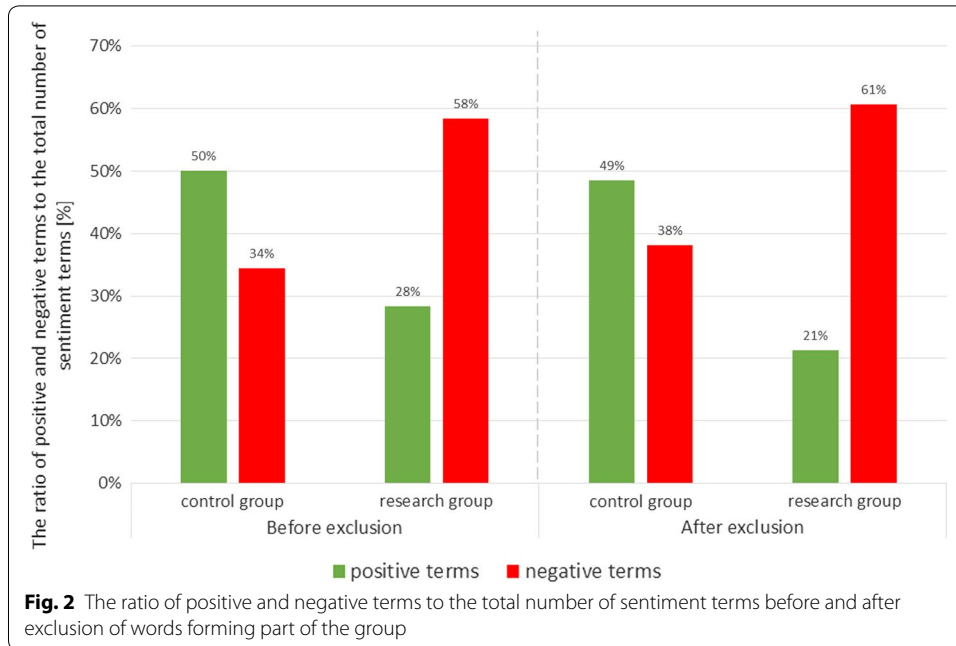
The analyzed texts are subjective and their form remains open for authors. Each document was saved in a separate text file. People who participated in the study have given their written consent.

### Results

SAS Institute Company tools were used for data processing, and in particular: SAS Text Miner [16] and SAS Sentiment Analysis [17]. The first tool was used to preprocess the texts to create a bag of words representation. In the next stage, sentiment analysis studio was used to analyze the documents according to the dictionary. As regards basic statistics regarding words taken from the documents in research group 15 unprocessed documents consist of 1146 words. After processing 500 words in the basic form remained. The average length of the documents was 76 words, 120 sentiment keywords were founded—an average of eight in the document.

Figure 1 shows the number of documents in research and control group which were classified in particular polarization groups (positive, neutral, negative, and undefined). The classification was based on the rules presented in “Sentiment analysis” section. A deeper analysis of the results obtained the sentimental media found in the documents. This was further analyzed, separating the common terms that occurred in the two study groups. Figure 2 shows the results of this analysis a comparison of the number of keywords found before and after excluding words occurring in both groups. By comparing the results for the data before and after excluding the common parts, the results show





that the general trend has not changed. For both sets of groups there is a comparable relationship: texts by healthy people contain more positive terms than negative ones, while texts by sick people contain more negative terms than positive ones. Below a summary of key words found in individual documents in the variant processed into the bag of words form (Table 1). Words repeated within one document have been reduced to one occurrence. Designated keywords were terms derived from the sentiment dictionary used in the work.

**Table 1** List of key words found in individual documents for the research group

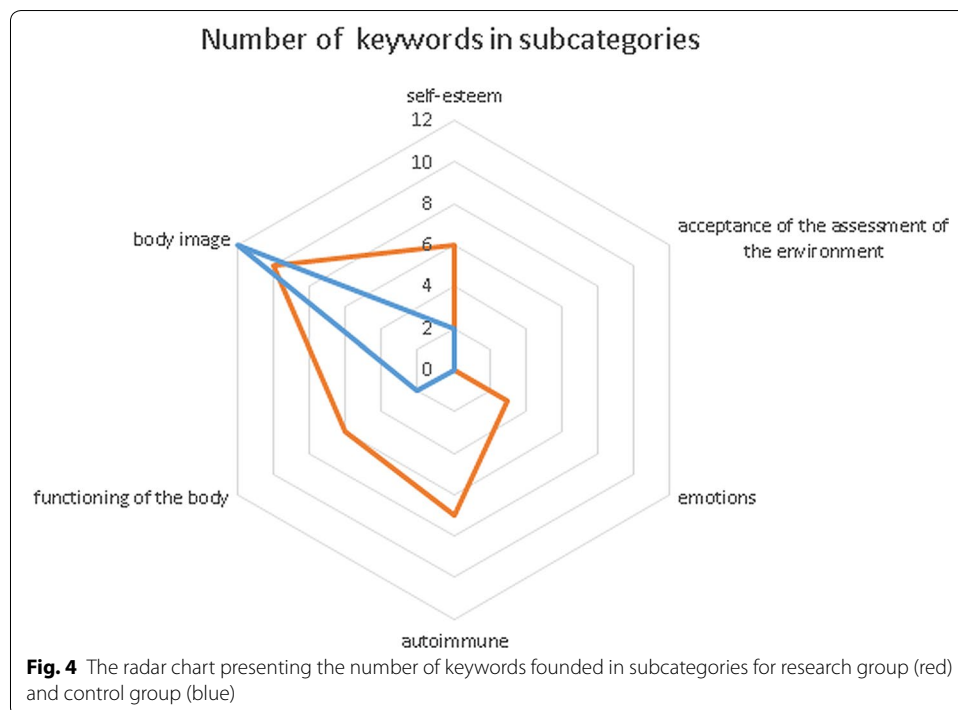
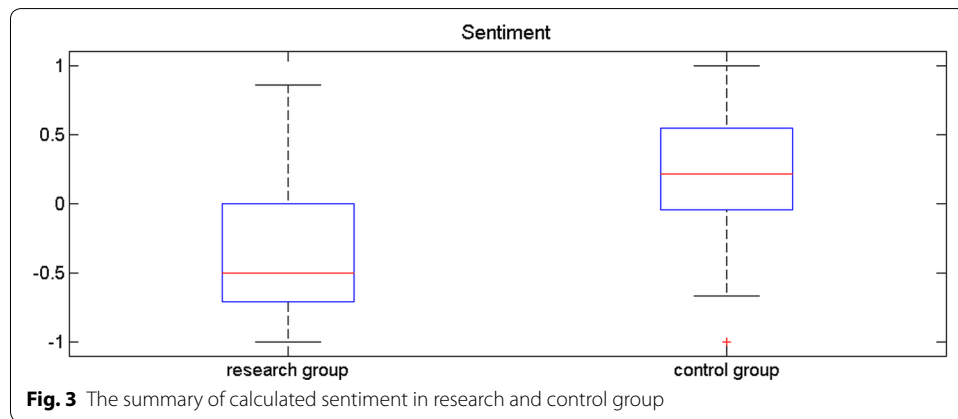
Id	Positive keywords	Negative keywords
1	Well, happy, satisfaction	Bone, skinny
2	Accept	Stupid, boring, ugly, sad, fat, ugly, horrible, hate, the enemy
3	Ideal, happy, easy, please, sure	The disease, worse, long, fat, obese, losing weight, bone, terribly hard, lose, shred, fatty, exaggeration, the lack of
4	A simple	Ugly, disproportionate
5	The impression	Fat, unattractive, obese
6	-	Wine, to punish, hurt themselves
7	Like, nice	Worse, embarrassing
8	-	Uptight
9	The joy of	Weak
10	Cool, enjoy	-
11	Advanced	The disease
12	-	Last, nobody, control, huge
13	-	Imperious, vomiting
14	Enjoy	Pretend nothing, losing weight
15	Accept, an amazing, satisfaction, strong, tenacious, help, well	Problem, fat, little

Figure 3 presents the summary of calculated sentiment. The calculated sentiment for research group is statistically significantly different from the sentiment for control group (two-side Wilcoxon rank sum test with 0.05 significance level—p-value: 0.0025).

To make the results more useful, the specific subcategories were introduced. These subcategories make it possible to present quantitative analysis of the results by showing the dictionaries and statistics of used words for specific person in every subcategory. After this consideration the normalized sentiment values could be present in a radar chart. Figure 4 presents the number of keywords founded in subcategories.

### Discussion

The described analysis indicates the existence of a relationship between the mental state of the author’s textual health and the vocabulary he or she uses. It is possible to indicate a set of characteristic sentiment terms specific to a given group of persons. Their





presence is related to the author's mental state and their body image. We analyzed the documents which were classified as negative in control group. Almost of them include criticism about the body. It seems, that for a large group of teenage girls, despite the fact that they do not show eating disorders, it is a big challenge to accept their bodies. It is a fact that also gives authors feeling of sadness and a lot to think about the pattern of beauty and attitudes towards the body that are widespread in society. Taking into account our experience, we have developed sub-categories that cover issues that a therapist meets during diagnosis and treatment of anorexia. In current studies, no medical history was collected, covering general health, but in general it can be said that anorexia was the most annoying disease in the research group. In the future, we plan to conduct research on the vocabulary of patients suffering from terminal diseases, including cancer. Then we will make a quantitative comparison of the statistics of the words used.

To use the quantitative results for a specific patient, it is necessary to check if there are enough sentiment terms for specific subcategories. If the therapist thinks useful information is presented in such a chart (Fig. 3), it is possible to explore the data by analyzing the sentimental media used in the individual subcategories and their statistical summaries. In addition, if patients are available from different treatment periods, they can be compared. It is possible to define the minimum number of sentiment terms to make this analysis credible and useful. It should also be emphasized that writing is a big challenge for patients with anorexia. Another option is to record the patient's behavior after their consent. Then, speech-to-text technology can be used to automatically create transcripts of the patient's speech. This option would help gather more text and build a personalized dictionary of the patient.

Analysis of emotional language by automated tools has additional challenges due to the ambiguous nature of the text. This analysis is further complicated by the characteristics of the language itself, such as: the impact of syntax on the meaning of each word, ambiguity of terms, the existence of metaphors, irony, or idioms. There are also misspellings that appear in the statements, which in many cases prevent proper recognition of the used words [15]. In addition, it is necessary to classify sentiment information to the particular topic. Most of the problems in the presented approach were eliminated by introduction of detailed subcategories, spelling correction, and word reduction to their basic forms.

Generally, it should also be taken into account that the texts of different authors are characterized by varying levels of subjectivity and intensity of emotion related to their personality [2, 10]. These factors will be included in future research by introducing sentiment weight, or an importance of specific terms for each person. A prerequisite for this improvement will be to increase the amount of text available for each person.

## Conclusions

Diagnosing and treating anorexia nervosa is an important challenge for modern psychiatry. Taking into account a connection between the mental state of a person and the characteristics of their language, this paper presents developed and tested method for analyzing the written statements of patients with anorexia nervosa and healthy individuals, including the identification of keywords.

The analysis was performed on the speeches of 67 people with the subject “My body”. These statements were made in written form and divided into two groups: research group—statements of ill persons (15 documents) and control group—statements of healthy people (52 documents). Due to the short nature of the texts, which is related to the difficulty of expressing oneself about one’s body when suffering from anorexia, the bag of words approach was used. In the first step (analysis of documents using a dictionary), a rule-based model was created. As a collection of rules, dictionary files were used corresponding to three groups of positive, negative, and neutral sounds. Next the analyzed texts were searched and counted keywords. Based on the keywords found, each of the documents was categorized into one of the groups in every subcategory. If the number of positive terms is greater, equal or less than the number of negative terms, the document is positive, neutral or negative. Where no keyword was found, the sentiment of the document is indefinite.

It is possible to indicate a set of characteristics sentiment for every person. Additionally, the results of specific patient could be analyzed in six specific subcategories: self-esteem, acceptance of the assessment of the environment, emotions, autoimmune, functioning of the body and body image. It could help focus on specific topics during therapy.

To use the quantitative results for a specific patient, it is necessary to check if there are enough sentiment terms for specific subcategories. It is possible to define the minimum number of sentiment terms to make this analysis credible and useful. It should also be emphasized that writing is a big challenge for patients with anorexia. Another option is to record the patient’s behavior after their consent. Then, speech-to-text technology can be used to automatically create transcripts of the patient’s speech. This option would help gather more text and build a personalized dictionary of the patient.

#### **Authors’ contributions**

Conception and design of the work: DS. Data analysis: KN. Psychological background, analysis and description, data collection: KR. Drafting the article: DS, KR, KN. Final approval of the article: DS. All authors read and approved the final manuscript.

#### **Author details**

<sup>1</sup> Faculty of Biomedical Engineering, Silesian University of Technology, Roosevelta 40, Zabrze, Poland. <sup>2</sup> Clinical Hospital No. 1 in Zabrze, Department of Pediatrics, School of Medicine with the Division of Dentistry in Zabrze, Medical University of Silesia in Katowice, 15-18 3th Maja, 41-800 Zabrze, Poland.

#### **Acknowledgements**

We want to thanks Andre Woloshuk for English language corrections.

#### **Competing interests**

The authors declare that they have no competing interests.

#### **Availability of data and materials**

The data that support the findings of this study are available from corresponding author in statistical processed representation in the numerical form of columns of the bag of words matrix.

#### **Consent for publication**

Not applicable.

#### **Ethics approval and consent to participate**

Bioetic Commission of the Silesian Medical University—approval is not needed. We used only anonymised notices of psychologist and patients who participated in the study, have given their written consent.

#### **Funding**

This research was financial supported by the Directional Research at the Faculty of Biomedical Engineering at the Silesian University of Technology in Gliwice, Poland BK-209/RIB1/2018.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 9 December 2017 Accepted: 27 January 2018

Published online: 02 February 2018

## References

1. Kaye W, Fudge J, Paulus M. New insights into symptoms and neurocircuit function of anorexia nervosa. *Nat Rev Neurosci*. 2009;10:537–84.
2. Delinsky S. Body image and anorexia nervosa. In: Cash T, Smolak L, editors. *Body image: a handbook of science, practice, and prevention*. 2nd ed. New York: Guilford Press; 2011. p. 279–87.
3. Izydorczyk B, Rybicka Klimczyk A. Diagnoza psychologiczna poznawczych i emocjonalnych aspektów obrazu ciała u dziewcząt i młodych kobiet polskich. *Problemy Medycy Rodzinnej*. 2008;10(4):24–35.
4. Sarol Kulka A, Kulka Z. Obraz własnego ciała a seksualność kobiet z zaburzeniami odżywiania się. *Post Psychol Neurol*. 2005;14(2):115–22.
5. Brytek Matera A. Postawy wobec ciała a obraz samych siebie u kobiet z zaburzeniami odżywiania się. Badania na gruncie teorii rozbieżności ja edwarda tory higginsa. *Psychiatria Polska*. 2011;45(5):671–82.
6. Jagielska G, Wolanczyk T, Osuch B. Zaburzenia miesiaczkowania w jadłowstrecie psychicznym. *Psychiatria Polska*. 2010;44(2):277–86.
7. Falk J, Halmi K. Amenorrhea in anorexia nervosa: examination of the critical body weight hypothesis. *Biol Psychiatry*. 1982;17(7):799–806.
8. Rojek K, Opoczynska Mosierowicz M. Ciało, które znaczy i naznacza. analiza fenomenologiczna doświadczania ciała przez osoby chorujące na anoreksję. *Psychoterapia*. 2014;1(168):35–46.
9. Rohnka N, Szymczyk B, Rusanowska M, Holas P, Krejtz I, Nezelek J. Właściwości języka osób cierpiących na zaburzenia emocjonalne i osobowości—analiza treści opisów codziennych wydarzeń. *Psychiatria Psychoterapia*. 2015;11:3–20.
10. Lyons E, Mehl M, Pennebaker J. Pro-anorexics and recovering anorexics differ in their linguistic internet self-presentation. *J Psychosoma Res*. 2006;60:253–6.
11. Rojek K, Opoczynska-Morasiewicz M. Ciało, które znaczy i naznacza. analiza fenomenologiczna doświadczania ciała przez osoby chorujące na anoreksję. *Psychoterapia*. 2014;168(1):35–46.
12. Ali T, Schramm D, Sokolova M, Ilkpen D. Can i hear you? Sentiment analysis on medical forums. In: *Proceedings of international joint conference on natural language processing*, 14–18 October 2013, Nagoya; 1996. p. 667–73.
13. Wilson T, Wiebe J, Hoffmann P. Can i hear you? Sentiment analysis on medical forums. In: *Proceeding HLT '05 proceedings of the conference on human language technology and empirical methods in natural language processing*, 06–08 October 2005, Vancouver; 2005. p. 347–54.
14. Amine A, Elberrichi Z, Simonet M. Evaluation of text clustering methods using wordnet. *Int Arab J Inf Technol*. 2010;7(4):349–57.
15. Pang B, Lee L. Opinion mining and sentiment analysis. *Found Trends Inf Retr*. 2008;2(1–2):1–135.
16. SAS Text Miner. 2017. [https://www.sas.com/en\\_us/software/text-miner.html](https://www.sas.com/en_us/software/text-miner.html). Accessed 9 Dec 2017.
17. SAS Sentiment Analysis. 2017. <http://support.sas.com/software/products/san/index.html>. Accessed 9 Dec 2017.

Submit your next manuscript to BioMed Central  
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

