Psychosocial Factor Identification in Cancer Patient Community

Ibrahim Moamen

University of Oulu

Mourad Oussalah (✉ Mourad.Oussalah@oulu.fi)

Oulun Yliopisto <https://orcid.org/0000-0002-4422-8723>

Dan Ackers

Oulun Seudun Ammattikorkeakoulu

Ekaterina Gilman

Oulun Yliopisto

Jouko Miettunen

Oulun Yliopisto

Research article

Keywords: Text mining, ontology, health, psychosocial factors

Posted Date: August 3rd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-41551/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

The open Suomi24 discussion forum, Finland's largest topic-centric social media, provides excellent opportunities to study users' interactions with various contemporary discussion topics. Its easy access and anonymity enable patient and users to seek valuable advices or share their experiences. This paper focuses primarily on cancer health users' discussion forum and attempts to lie down foundations for content-analysis based approach for identifying psychosocial factors of patient cancer community in Suomi24.

Method

The methodology utilizes to a large extent innovative natural language processing based approach that calls for ontology based construction for identifying the relevant discussion threads in Suomi24 corpus. Next, a co-occurrence analysis with various distance level was employed to quantify the strength of the co-occurring words. Inspired by psychosocial factors analysis in related medical studies, a categorization for psychosocial factors has been put forward and investigated.

Results

The results overlap to a large extent with some previous findings in clinical studies where the dominance of the social factors is highly reported. An open platform has been put forward to help both researchers and clinicians to study the Suomi24 collected dataset, with the potential correlation analysis with patient records through dedicated electronic medical record databases.

Conclusion

This work paves the way for further investigation of medical textual data in order to correlate the finding with patient records as registered in electronic medical records. Besides, the analysis also highlights the challenge of accommodating standard NLP tools to Finnish language where more work is needed in order to match the performance of the available tools in English textual analysis.

Background

The exponential increase of social media across all population groups and disciplines paves the way for new research directions in behavioral science, information technology, and health science. In the latter, social media has been widely acknowledged by its ability to convey social support to patients and their relatives/friends. For instance, social media was used for sentiment analysis [1], comprehending patient perception of a particular topic by patients [2], mining social emotion from textual content [3], mapping evidence related to suicide [4] and depression [5-7], feelings [8], among others. In this respect, Cohen et al. [9] reported that the social support is primarily manifested through four distinct categories: i) emotional support, which helps patients to meet their emotional or affective needs; ii) esteem support that boosts the patient's confidence to overcome inherent challenge(s); iii) information support, which provides the patient with unforeseen or useful tricks from other experiences and community; iv) network support that reminds the patient about the potential helps he/she can get from the community he/she belongs to. These findings have been identified by surveying the content of the main social media platforms; namely, Facebook, Instagram, Twitter, and SnapChat.

Suomi24 [10] is Finland's largest topic-centric social media as well as one of the largest non-English online discussion forums in the world [11]. According to TNS metrix (2013)^[1], every month, 86% of Finns using the internet visit the site – at least to find out an answer to a question they have googled. The popularity of the forum arises from the anonymity of the conversations, which motivates the users to engage freely in various discussion threads that include political trends, policies, health issues, news, and hobbies. This opens up opportunities for methodological development, data analysis and visualization tools in order to characterize the posts as well as their contextual, social, and temporal dynamics. More

specifically, Suomi24 dataset covers two decades of online discussion history and contains over a 2B token corpus of Finnish discussion with almost 832.000 unique users per week each of whom on average spends 7 minutes on the forum at a time. This makes Suomi24 dataset an excellent laboratory for exploring the combining linguistic characteristics of text and statistical features of the discussion dynamics with social science research questions, opening up unique perspectives for combining qualitative trends with quantitative dynamics of interaction, users' emotions and sentiment, or explore various facets of conflicting evidence. Academically, this offers a clear advantage to other social media, such as Facebook and Twitter, whose complete set of data is normally unattainable for such critical examination. This motivates the current use of the dataset for health-like analysis exploring the psychosocial factors of cancer patients.

Anonymous online discussions are known to provide an important arena for addressing and sharing health-related problems [12-14] where treatment efficiency, side effects, and personal experiences are very popular. Indeed, medicines and their dosages, side effects and availability are referred to, for example, when discussing the diagnosis and progress of diseases. Besides, the criticism of health care practices is often associated with the limited, inadequate or excessive medication of people and overt medicalization of people's lives. Overall, anonymous social media data open up a perspective to peer-to-peer talk on what might be thought of as taboos in health issues, bypassing the standard professional setting of doctor-patient relation. From this perspective, talk about medication and associated lifestyle can be seen as part of the arena where people are conceptualizing and negotiating their health concerns together with the suitability of current practices and constraints. Life-threatening cancer patients are often diagnosed by anxiety and depression, which require special attention from the community as a whole, see, for instance, [15]. Forum is often an excellent venue for conveying and highlighting such support where patients also attempt to broaden their cancer-knowledge and validate /refute information received from health care professionals. This extensively occurred in Suomi24 discussion threads as per prior investigations [16]. This includes emotional support through various expressions of care and concerns that ultimately help to improve patient's mood, sharing of experiences and new routes that may raise patient awareness, esteem and social network in a way to boost confidence while fostering (new) relationship based on shared attributes. Loosely speaking, this also contributes, to some extent, to the empowerment of the online cancer patient community.

Research on blogs ranges from blog identification [17], identifying the source of re-publication of blogs [18], link analysis [19], buzz analysis [20] and social network analysis [21]. A number of tools have also been developed to aid in visualizing all sorts of blog data, mainly with a focus on visualizing the connections between blogs and individual bloggers, or the dynamic of buzz terms, topics or selected attributes over time, causation content of blogs, among others, see, e.g., [22].

Although there is a long history of analyzing, visualizing and exploring English online discussions with text mining and machine learning methods [23], the effective handling of Suomi24 dataset is far to be an easy task and several challenges were identified. First, although the forum promotes thread-based discussion, it often occurs that health or cancer-related conversation takes place in non-health-related threads. Second, the anonymity and the absence of the concept of user ID renders any network like analysis quite difficult and limited. Third, the use of standard natural language processing tools is quite limited for the Finnish language. Indeed, whereas in English the vocabulary size tends towards a log-curve as the size of the material grows, in Finnish it is typical that as the size of the data set grows, the number of word forms continues to grow linearly [24]. This leads to data scarcity, which is a problem for many machine learning methods that rely for instance on bag-of-words like representations of data. Moreover, in colloquial texts, rife with misspellings, acronyms, and invented words, the usefulness of linguistic analysis is limited. Fourth, the use of open-source translation API often yields unsatisfactory responses due to its inherent limitation in both scope and constructs. Fifth, cancer is one of the most sensitive issues that patients generally do not feel comfortable writing about in detail, which adds extra difficulty to researchers to detect useful clues, and potentially identify appropriate metrics relevant to clinicians.

The current study proposes a practical tool that can be employed to explore the Suomi24 dataset to identify potential patterns that are associated with health queries. Especially psychosocial factors of cancer patients are investigated. Therefore, the main research questions highlighted in this article are 1) How to identify the psychosocial factors from online discussion forums? 2) How to design an automated approach that identifies such factors in Suomi24?

The solution proposed in this paper tackles the above research questions restricted to cancer patient discourse using a combination of natural language processing techniques, statistical analysis and sound statistical observations. Especially, original methodologies have been devised and implemented in order to (i) collect relevant dataset from overall Suomi24 corpus, (ii) identify the main psychosocial factor categories, (iii) characterize each category through a specific ontology, (iv) evaluate the extent of each psychosocial factor in the corpus, and (v) interact with a wider community through a dedicated portal. Nevertheless, we believe that such a developed approach can ultimately be employed to explore other socio-linguistics patterns from the Suomi24 dataset. The rest of the article is structured as follows. Section 2 of this paper details the data collection procedure, key features of the dataset and the preprocessing methodology. Section 3 deals with the data analysis, which includes the corpus analysis and identification of psychosocial factors. Section 4 highlights the platform implementation. Finally, we conclude the article and highlight future work.

[1] <http://tnsmetrix.tns-gallup.fi/public/>

Methods

Suomi24 and healthcare data

Suomi24 is a discussion forum that consists of publicly-available, user-generated discussions grouped by content, such as entertainment, hobbies, travel, and health. Being anonymous, users can start their discussions or contribute to existing ones. In this study, the forum data was retrieved from a structured database, accommodated by the service provider. The license of the database, in compliance with copyright agreements by the World Intellectual Property Organization, grants the right to use and make copies of the corpus for educational, teaching and research purposes [25]. The dataset contains 352,725 posts in the Health category divided into 16 sub-categories, namely "ask your health questions", "birth control", "decease and mourning", "diseases", "drugs and addictions", "general health", "healthcare", "healthcare services", "medicines", "men's health", "mental health and wellbeing", "oral health", "plastic surgery", "senses (sensory organs)", "weight control", and "women's health". The distribution of the number of observations among the 16 categories is not uniform, see Table 1. Titles and posts contain 2.6 and 75.9 words on average, respectively. The median values of word counts are 2 for titles and 49 for posts as highlighted in Table 1. On the other hand, we also tried to quantify the extent to which the posts contain unknown wordings. The latter is obtained by looking into the presence of individual word at Finnish dictionary or one of the Finnish SMS databases (e.g., Slangi.net, Suomen murteiden sanakirja). Therefore, if a word can be found neither in the Finnish dictionary nor in the two SMS databases mentioned above, it is classified as Unknown in Table 1.

The review of health-related discussion threads in Suomi24 shown in Table 1 reveals the following:

- There is a surge in discussions related to diseases and mental health/wellbeing as indicated by a large number of posts in these threads.
- The users use lengthy writing when it comes to mental health issues, possibly to show their emotions and feeling while attempting to pay special attention to patient experience and argumentation. It also indicates the willingness to share experiences.
- The occurrence of unknown wording, although small, is found to be relatively more important in Diseases thread and Medicine thread than others. One of the explanation for this observation arises from the technical terminology employed for diseases and medicine, which combined with users' voluntary/non-voluntary misspelling errors, resulted in unknown wordings.
- Despite its importance, "cancer" is not mentioned as a separate thread in the health discussion forum. This renders the search for cancer-related discussion spread across multiple threads. For instance, cancer-related wordings are found in

Diseases, General Health, Healthcare, Mental Health & Wellbeing threads. This trivially does not mean that cancer-related wording cannot be found elsewhere.

TABLE I. Suomi24 Health dataset statistics

Category	#posts	Word Counts		
		Mean	Median	%U words
Ask your health questions	6,012	74.5	57	2.6
Birth control	13,213	58.6	47	2.4
Decease and mourning	4,108	94.3	62	2.9
Diseases	86,035	76.4	52	4.8
Drugs and addictions	34,346	68.6	35	2.8
General health	16,415	69.2	49	2.1
Healthcare	15,461	61.0	34	1.8
Healthcare services	236	61.5	35	1.7
Medicines	9,743	52.5	36	3.5
Men's health	4,067	55.8	36	2.3
Mental health & wellbeing	70,017	103.0	66	1.8
Oral health	10,959	59.3	43	1.9
Plastic surgery	6,123	55.6	39	3.7
Sensory organs	8,164	64.5	46	2.1
Weight control	49,257	69.1	47	1.6
Women's health	18,569	69.5	55	1.7

The absence of thread related discussion highlights the importance of keyword related search across the whole health Suomi24 dataset to identify key patterns that might be associated with cancer patients.

Data Collection

The first step in our methodology is to gather cancer-related dataset from Suomi24 forum. The forum provides an API to get forum posts from the database. HTTP POST request with a "keyword" parameter was used to retrieve the corresponding discussions. For this purpose, we select a bag of keywords that include:

- Cancer* and its synonyms as gathered from online Finnish dictionary,
- Slang terms employed for cancer and cancer-related diseases gathered from slangui.net
- Medical dictionary of cancer-related diseases extracted from Bupa.co.uk (www.bupa.co.uk/health-information/cancer/cancer-terminology) together with their Finnish translation using Google translation API,
- The NCI Dictionary of Cancer terms (8,334 terms) from NIH National Cancer Institute, which covers drugs, cancer types and its various clinical (universal terminology, so no translation needed),
- Finnish cancer glossary created by Prof. Teppo Lyly, available at <https://www.kaikkisyovasta.fi/tietoa-syovasta/syopasanasto/>

The format of the returned dataset is JSON, whereas Python library “ijson” was used to loop through all the available data. Therefore, the procedure to select the research dataset is as follows: First, the entire health-related dataset pointed out in Sect. 2.1 is taken into account and considered as one source for cancer patient discussion. Second, we run a text search across the whole Suomi24 dataset for the above bag of keywords, and whenever a match of at least one keyword is found, the associated post, name of thread, five posts prior the matching post as well as the following five posts (provided the post does not occur at subsequent thread) are retrieved. The desire to retrieve prior and subsequent posts of the matching post together with the underlying thread are motivated by the concern to capture the context of the discussion. The overall schema is presented in Fig. 1.

Data Preprocessing

The preprocessing task aims to identify the boundaries of individual posts and their various constituents (tokens) taking into account a large number of potentially misleading content and writing styles that may cause unsatisfactory segmentation. This includes, for instance, the presence of links, uncommon characters, miss-placed spaces or punctuation characters, which can impact further analysis. For this purpose, for each post in the cancer dataset, the following tasks have been performed.

- i. All text is converted to lowercase letters/characters;
- ii. Whitespace is added after each dot unless if not available;
- iii. Any number of consecutive whitespace characters are transformed into a single whitespace character;
- iv. All URLs are removed together with any uncommon character (e.g., Chinese and Greek characters). Odd characters are also removed using regex-functions;
- v. Stopwords are removed to reduce the list of keywords. This list is borrowed from UniNE of Neuchâtel University, after manual checking to keep all adjectives, time/date information and negation related terms due to their importance in subsequent reasoning;
- vi. Tokenization is carried out to identify individual tokens of the posts. The classical white-space delimiter cannot work in the Finnish language due to the inherent structure of language that combines individual words to create more lengthy words. Therefore, the contribution of the Finnish language parser is ultimately important to distinguish individual wording. For this purpose, we use Finnish-deep-parser developed at Turku University;
- vii. A simple stemming approach is adopted. This aims to identify the root/stem of a word to reduce the number of word inflections, which, in turn, reduces the total number of keywords/tokens used to represent the sentence (post). For this purpose, we employed FinnPos an open-source morphological tagging and lemmatization toolkit for Finnish.

Data Analysis

Initial Analysis

To comprehend the content of the gathered dataset, we report the distribution of the most frequent words in the filtered corpus and the frequency of the most occurring cancer terms to visualize the closeness/farness with the most occurring terms. Algorithm 1 described in Fig. 1 was used for term-frequency analysis. A summary of the attributes of the dataset is described in Table 2.

TABLE II. SUMMARY OF DATASET

Size of dataset	400MB
Number of posts	55500
Number of threads	26052
Number of distinct words	41024
Time frame	2001-2015

The results are presented in Table 3.

TABLE III. MOST OCCURRING TERMS & CANCER RELATED TERMS

Most repeated words		Most repeated cancer words	
Word	#number	Word	#number
saa	21862	määrä	3190
mun	7190	isä	1718
mies	7123	kuoli	1447
ihminen	6623	leikkaus	1324
apua	6505	raha	1197
euroa	6409	hoidon	1189
kokemuksia	5886	tauti	1039
tietää	5822	kasvain	345
päivä	5807	kaveri	248
oon	5793	tutkimus	179
nainen	5755	lääkitys	135
osaa	5311	hoito	107

Commenting on the finding, one should notice the following:

- The results were obtained after removing the stopwords, which would discard less useful terms in the study.
- For the clarity of illustration, we restricted to the most twelve frequent terms in each category (non-cancer-related terms and cancer-related terms). This is also motivated by the fact that the above terms constitute almost half of the terms which have at least three occurrences in the identified sample.
- The majority of non-cancer terms are related to affection, and various types of help (sharing experience, finance, advice).
- The graph also highlights the quasi-dominance of multi-form ("saa"), which translates the action of "getting", due to its wide usage in Finnish.
- The cancer-related terms are as expected less frequent than non-cancer-related terms by a factor of at least two. Besides, the treatment-related terms are even less frequent, which is motivated by the context employed to generate discussions around cancer-related terms.
- The cancer discourse highlights the importance of financial aspects, treatment effects, death, disease, and their medication.

Evidence mapping

Co-occurrence analysis

The first stage in evidence mapping is to explore the most frequent co-occurrences in the filtered dataset as an indication of relatedness between the terms. In other words, this hypothesizes that the higher the co-occurrence between two terms, the higher the association between these terms.

Further development incorporated the functionality to distinguish the extent to which the co-occurring terms felt apart from each other. This reflects on the hypothesis that co-occurring terms that occur close to each other in the sentence or post are deemed more important than those terms that co-occur far apart from each other. On the other hand, to take into account the syntactic relationship among the co-occurring terms, a dependency tree was used; namely, Turku Dependency Treebank was employed to deal with the Finnish language. An example of the use of the Turku Dependency tree is shown in Fig. 3.

The use of a parser tree allows us to capture the relationship between wordings even if they are located far apart from each other. For instance, a verb and its subject, as well as its complement, would entail a direct link from the verb to its subject and from a verb and its complement. Therefore, we hypothesize that whenever two words, say, w_1 and w_2 are directly linked to each other in the view of the parser Tree, then the distance between these two words is equal to one even if they are far apart from each other, otherwise, the distance is calculated as the number of tokens that separate w_1 and w_2 . More formally, this can be formulated as follows

$$d(w_1, w_2) = \begin{cases} 1 & \text{if } w_1 \text{ and } w_2 \text{ are directly linked in the parser} \\ \text{number of tokens between } w_1 \text{ and } w_2, & \text{otherwise} \end{cases} \quad (1)$$

From an implementation perspective, we restricted to five different distance levels: 1–2 word distance, meaning next to each other (either directly or through dependency tree structure), which triggers a strong influence on each other; 3–5 words distance, indicating potentially significant impact on each other and; 6–7 words apart, where the impact is likely to be medium. Finally, levels 9–14 words (apart from each other) translate a rather minor to low impact on each other, and 15 + distance (more than 15 words apart) implies that the co-occurrence is considered to have no impact on the syntactic, semantic or causal relationship between the two terms w_1 and w_2 . Algorithm 2 (see Fig. 4) highlights the implementation of the distance-based co-occurrence paradigm. Figure 5 highlights the word co-occurrences taking into account the various distances. While Table 4 reports on the top co-occurring pairs corresponding to general discussion and cancer related terms.

TABLE IV. OVERALL WORD CO-OCCURRING OF CANCER AND NON-CANCER RELATED TERMS

Most repeated words		Most repeated cancer words	
Word	#number	Word	#number
jostain-syystä	1709	maksaa-euroa	251
pari-päivää	1687	hoidot-tietoa	102
pari-viikkoa	1683	hoidot-hoitohen	101
osaa-sanaa	1632	hoidot-tehokkuudesta	101
viikon-päästä	1568	hoidot-varmaa	101
seuraavana-päivänä	1562	äiti-isä	101
päästä-eroon	1474	hoitoa-olemassaolevaa	97
lääkäri-sanoi	1466	hoitoa-tutkimusnäyttöä	97
pitäisi-tehdä	1427	hoidot-Tutkijaryhmän	96
the-and	1411	hoidot-jaottelemisen	96
ottaa-yhteyttä	1186	hoidot-kategorioihin	96
and-the	1168	hoidot-myöntää	96

A quick reading of the results reveals the following:

- The scrutinizing of the highly co-occurring words, outside cancer-related discourse, shows the dominance of explanatory proposition as in (jostain-syystä), which translates (for some reasons) and timing propositions as in (pari-viikkoa), (viikon-päästä), which are translated into (couple of weeks), (in a week), respectively. Conditional propositions can also be noticed in (pitäisi-tehdä) translating (would have to – make). Recommendation-like propositions are highlighted in the pair (ottaa - yhteyttä), translating to (take - contact).
- Despite the quasi dominance of Finnish language, some English terminology is not There exist many English wording, which justifies the dominance of the determinant expression (the) and conjunction expression (and).
- There is a single co-occurrence that is more related to clinical setting (lääkäri – sanoi), a translation of (a doctor – said), which indicates the high interest of the community users to seek doctor advice and recommendations from other experience perspectives.
- The scrutinizing of the cancer-related discourse indicates the dominance of the financial aspect where users/patients seek for treatment costs either directly or indirectly as in (maksaa-eur). The next co-occurring wordings are related to the type and effectiveness of the treatment (see, all pairs with “hoidot”). One also notices the emergence of family-related issues in (äiti-isä), translating (mother-father) relationship.
- The preceding highlights the importance of treatment effects, financial cost and family issues as dominant factors in the cancer discourse. Whereas the non-cancer discourse shows the importance of the clinical advice in Suomi24 health community, seeking doctors’ saying and confirmation. On the other hand, the dominance of conditional and recommendation expressions indicates the importance of the argumentation structures and storytelling scenarios where users attempt to build arguments to reply to patient/user worry or through sharing a personal experience.

- vi. The analysis of the results of Fig. 6 shows the large disparity of frequency values between the terms co-occurring at distance 1 (or occasionally, distance 2) and those co-occurring at higher distances (3–5, 6–8, 9–14, 15+) by an order of magnitude of more than 100! This indicates that the co-occurrences occurring at a distance of more than 2 are statistically insignificant when compared to those occurring at distance 1 or 2.

Psychosocial factor categorization

Rationality

In this study, the psychosocial factors are categorized into five categories: *Financial, Illness, Treatment, Social and Death*. The motivation for such categorization is based on commonsense reasoning, literature study and manual analysis of data. Indeed, intuitively, cancer patients foresee death as the ultimate end either in short or medium term. Besides, the type of illness, the effectiveness of treatment, its associated cost (both clinical cost and family-related cost) as well as patient social life are naturally of paramount importance to any cancer patient. Second, previous studies reported the clinical and psychosocial factors associated with anxiety and depressive disorders in breast cancer patients and other cancer patients [26]. The authors identified three types of associated psychosocial factors: social support, family relationship, and conflict solving. Anxiety and depression in breast cancer patients were usually common in environments of poor family relationships, which also contributed to more pain and fatigue as well [27]. This stresses the importance of social factors as part of psychosocial factors for cancer patients. Studies in [28] suggested more social support was associated with lower cancer incidence and mortality, and longer disease-free interval and survival, which triggers the importance of treatment, illness and death factors. Similarly, authors in [29] found that the number of social ties and experienced social isolation was related to cancer incidence and mortality from cancer. Mofatt et al. [30] have found that the onset, treatment, and trajectory of cancer are associated with the financial stress among patients, which has been identified as a significant unmet need. This stresses the importance of financial factors as part of the overall umbrella of psychosocial factors. Third, manual scrutinizing of many Suomi24 posts revealed that potential treatment success and failure (death) appear among hot discussion topics. Similarly, illness and treatment are usually associated with cancer conversations, where people usually seek education, peer support or advice to cope up with their disease and its associated challenges [31]. Time costs and out-of-pocket expenditures demonstrate the sizeable financial burden that patients usually endure, often, with limited resources, relying on family members or friends [32], which promotes the role of financial issues as part of the main psychosocial factors. Fourth, the term frequency and co-occurrence analysis carried out in the previous section undoubtedly supports the categorization above of the psychosocial factors. Indeed, a simple term frequency of cancer discourse reveals the dominance of financial aspects, treatment effects, death and disease (or illness) related issues. Whereas the non-cancer discourse highlights the dominance of social factors. Indeed, family relationships, for instance, were among the top co-occurring words in Fig. 5. Besides, the abundance of argumentation and storytelling expressions, which helps to share story-telling and build convincing statements in a way to answer patient worries, testifies on the significance of psychological/social support.

Ontology characterization of categories

Inspired by techniques employed in some hate-speech detection research (see, e.g., review paper of Schmidt and Wiegand [33]), the idea employed for building the ontology related to each category is the following. First, we start with a limited number of manually crafted words assigned to each category based on our understanding of the linguistic and contextual meaning of each category. Second, we perform a manual investigation of a random selection of around 1000 posts of the collected dataset. The aim of this investigation is to refine the manually crafted words of each category by amending or adding extra wordings that we found relevant to a given category after reading the post (s), where it often occurs that many identified terms through post review are rather less linguistically and semantically related to the actual meaning of the category. On the other hand, given that several homonyms words are resulting from the manual craft and random post scrutinization phase, we revisit each homonym word and assigns the corresponding sense (synset) that best matches the underlined category. We used FinnWordNet, provided by University of Helsinki as part of FIN-Clarn project (see, FinnWordNet [34]) to generate the associated synsets. This operation is independently performed by three Finnish native speakers in order

to eliminate any bias, conflict or dismiss of some interesting event (s). Third, using the (Finnish) WordNet lexical database, an automated system was created to add, for each word of a given category, its set of synonyms, antonyms, hyponyms, and hypernyms whenever available. If a newly constructed word (synonym, antonym, hyponym or hypernym) is found to belong to more than one category, it is systematically removed. Such reasoning although intuitive and innovative in its approach to generating a reasonably large collection of related words for each category has also its limitations as will be highlighted later on. The overall architecture of the ontology building is shown in Fig. 6.

Evaluation of categories in collected dataset

After generating the ontology associated with each psychosocial factor category, we would like to evaluate the occurrence of each category in the overall collected dataset.

For this purpose, since the ontology of each category is made of a set of terms generated through the process of Fig. 6, a simple string matching would enable us to determine the extent of the hit in the overall dataset. Nevertheless, special care should be devised for those homonym terms, which are represented by their synsets. In such cases, an extra step is taken to find out whether there is a matching of senses as well or not. Consequently, there is a need to have a mechanism in place to identify the correct synset of a given hyponym word in the post. For this purpose, a novel mechanism is designed that makes use of semantic similarity approach. More specifically, given a post P , tokenized as

$$P = \langle T_1, T_2, \dots, T_m \rangle$$

Assume, for instance, without loss of generality, that T_2 is the target word for which one seeks to determine the associated synset.

A blind approach that does not fully explore the whole syntactic relationship of the underlined sentence containing T_2 but is computationally appealing consists of (i) calculating the semantic similarity of T_2 with first term situated next to it for which such semantic similarity exists using, for instance, Wu and Palmer wordnet similarity (Sim_{WP}) [35], and then (ii) retrieving the synset of T_2 that yields such similarity score. In other words, this boils down to the following:

If $Sim_{WP}(T_1, T_2)$ exists then retrieves synset (T_2)

else If $Sim_{WP}(T_2, T_3)$ exists then retrieves synset (T_2)

else If $Sim_{WP}(T_2, T_4)$ exists then retrieves synset (T_2)

Etc.

The intuition behind such reasoning is that the wordnet semantic similarity score between two words T_i and T_j always outputs the maximum score that can be achieved when considering all combinations of synsets of T_i and T_j . Therefore, the generated synset corresponds to the sense that agrees most with the subsequent word in the post. This captures to a large extent the correct sense of the target word in the underlined sentence. The algorithmic implementation of the preceding is described in Algorithm 3.

Algorithm 3: Synset identification of homonym target word in the post

Input 1. homonym word W_k

2. Post sequence: $W_1, W_2, \dots, W_k, W_{k+1}, \dots, W_n$

Output Synset W_k^t

$t = k$

```

0: If  $\text{Sim}_{\text{WP}}(W_{k-1}, W_t)$  exists Return synset  $W_k^t$  of  $W_k$ 

Else If  $\text{Sim}_{\text{WP}}(W_t, W_{k+1})$  exists Return synset  $W_k^t$  of  $W_k$ 

Else  $k = k + 1$  GOTO 0

END

```

Next, once the process of matching the homonym words to the post is achieved, the overall evaluation turns into a simple string matching process as a proportion of the occurrence of each factor in the whole collected dataset. This is performed using a simple word-hit based methodology. More formally, given the i^{th} category of the psychosocial factor (Illness, Treatment, Social, Financial and Death), the associated proportion p_i is calculated as:

—

2

3

where c stands for a normalization factor, Post_j corresponds to the set of tokens associated to j^{th} post of the dataset, whereas C_i refers to all wording (tokens) that can be associated with the i^{th} psychosocial category. The sum in the above expression is performed over all the posts contained in the dataset.

Results

Psychological factor categorization

The strength of each of the five psychosocial factors using evaluations highlighted in expression (2–3) is illustrated in Fig. 7.

The results show the dominance of the social factors followed by financial and treatment-related issues. Intuitively, cancer patients experience many worries e.g. how their spouses or children manage after their onset of cancer, whereas financial burden may arise because of fear of salary loss due to illness, or potential salary cut. This is also positively correlated with other research findings. For instance, (McFarland and Holland [37], Falisi [38]) stress on the tendency of cancer patients to build a social life that overcomes the psychological effects of the treatment and illness, whereas the financial burden arises naturally as an implicit cost of either treatment or family-related charges.

Portal implementation

To enable a flexible solution that explores Suomi24 dataset, as well as providing sufficient user experience, we implemented a Web application supported by back-end database. As a platform, Apache HTTP Server was used as a web server, whereas Firebase was selected for the back-end database. The application was written mainly in JavaScript with React library. The use of web application allows easy access, and enables both clinician and academic researchers to test various research hypotheses and gain important insights regarding the behavior of the cancer patients as indicated by Suomi24 discussion forum. Figure 8 shows a user interface of the Web application with a query, where “Terveys”, which means “Health”, is one of the subtopics on Suomi24 discussions. The query result is shown in the same plot, where the co-occurrences and all the subtopics that happened to be in the detected discussions are presented, with a count of how many times it happened.

Visualization of effects of thread and psychosocial factors

Besides, an interface has also been designed in order to visualize, for each (highly frequently) co-occurring terms the effects of other metadata data of the Suomi24 dataset. This includes, for instance, the discussion threads (topics) where these terms co-occurred, the variation of word-distance between the co-occurred terms, and the psychosocial social factors identified across the various posts where the terms co-occurred. A snapshot from firebase with one example is illustrated in Fig. 9. Other functionalities include whether the co-occurring terms occurred in the title of the thread or sub-threads or only at the body of the post, the time slot and whether the username of the author is available.

Discussion

The above analysis although appealing due to its full autonomy and ability to accommodate expert knowledge in terms of ontology description of the various health-related concepts investigated throughout this study is not prone to inherent limitations.

First, as most data analytics techniques, the outcome is impacted by intrinsic limitations of text mining techniques. This includes (i) the effects of word normalization, which may eliminate some important elements of the sentence constructs and makes the output of the parser incorrect; (ii) the effects of composed words, which are sometimes not identified by the parser as so, trivially impact the performance of information retrieval task; (iii) the existence of multi-lingual expressions, where often English and Swedish expressions are found in Suomi24 dataset, which renders the parser output sometimes quite vulnerable.

Second, handling the Finnish language presents an extra difficulty in the sense that existing tools, including parsers, word normalization tools, and the FinnWordNet are still unstable as compared to English language tools. This trivially affects the accuracy of the results as well. Therefore, the output of the wordsense disambiguation of homonym words, for instance, is subject to the instability of the FinnWordNet and the subsequent calculus of semantic similarity score. Nevertheless, it should be noted that to a large extent the homonymy affects only a tiny fraction of the category ontology words, which restricts the scope of such limitation on the evaluation results.

Third, the use of initial manual crafting for initial design ontology associated with each category is trivially subject to the subjective judgment of authors in charge of manual labeling according to their understanding of the scope of each category. Besides, the inherent vagueness associated with the human language interpretation makes it very difficult to avoid overlapping wording that can naturally belong to more than one category.

Fourth, the recourse to a random selection of the posts to add / amend the initial crafting is subject to the randomness nature of those posts. Although the recourse to manual-post reading enables us to discover new jargon expressions employed by the users that would have been impossible to infer using solely simple linguistic interpretations of the scope of the categories, the extent of such jargon wording is far to be exhaustive. Indeed, many important wordings issued from local jargon describing an aspect of finance or social factors, for instance, might be overlooked. It is, therefore, not excluded that additional manual scrutinization of the posts would result in further expansion of the discourse of the category ontology. Nevertheless, we acknowledge that unlike microblogging, the extent of such jargon is not that abundant in the Suomi24. This is also in agreement with some other related studies involving the Suomi24 corpus as part of the Academy of Finland Citizen Mindscapes project [38].

Fifth, the result of the first and second phase in the ontology construction reveals the existence of several wordings belonging to the same root word but assigned to distinct categories. Accordingly, such wording would induce the same hyponyms, hypernyms, which would result in a cancelation of these words as part of the overall ontology construction. Likewise, regardless of whether the above words are of part of same word inflection or not, the use of hyponymy and hypernym relationship may also lead to words that are already assigned to other categories.

Sixth, the idea of systematically removing words that belong to distinct categories as part of design methodology can also be questioned. Indeed, although this expects to reduce the potential overlapping among these categories, which, in turn,

increases their distinguishability, it can lead to an underestimation of their occurrences. Loosely speaking, the effect of such underestimation is not important because the core aim is rather to comprehend the extent of variations that exist among the occurrences of these categories in the corpus instead of an accurate estimation of the occurrence of individual category.

Conclusion

This paper provides an open platform for analyzing Suomi24 health data in order to identify psychosocial factors associated to cancer patients. The methodology makes use of innovative natural language processing and data fusion techniques that calls for co-occurrence analysis with various distance level in a way to quantify the strength of the co-occurring words. On the other hand, inspired by psychosocial factors analysis in related medical studies, a categorization for psychosocial factors has been put forward and investigated. The results reinforce some of previous finding in clinical studies related to the dominance of the social factors. An open platform has been put forward to help both researchers and clinicians. The work paves the way for further investigation of medical textual data in order to correlate the finding with patient records as registered in electronic medical record. Besides, the analysis also highlights the challenge of accommodating standard NLP tools to Finnish language where more work is needed in order to match the performance of the tools in English text. Especially, the concept of textual argument and text summarization, in continuation of our work in [39–40], can be employed to ensure some inherent consistency and coherence to the textual outputs.

Declarations

Ethics approval and consent to participate

This study was approval by the institutional review board of the University of Oulu Hospital (UO15 2017-2019).

Consent for publication

Not applicable.

Availability of data and materials

Data generated or analyzed during this study are included in this published article and its supplementary information files with either appropriate links or excel data sheet.

Competing interests

The authors declare that they have no competing interests

Funding

This research is supported by funding from the Finnish Cancer Research Association #170129 2017/2019.

The views expressed in this publication are those of the author(s) and not Finnish Cancer Research Association. The funders had no input into the design of the study, the collection, analysis, and interpretation of data nor in writing the manuscript

Authors' contributions

MI and DA contributed to the software implementation of the platform and the statistical design of this study. MO was a major contributor in writing the manuscript and technical supervision of the software implementation and statistical analysis. EG contributed in technical supervision of DA work. JM contributed to the conceptualization and funding acquisition of this study as well as monitoring the ethical approval of this study. All authors read and approved the final manuscript

Acknowledgement

The authors would like to thank Cancer research clinician team from University of Oulu involved in the Cancer Project for their stimulating technical discussions during the project meeting, which helped the authors to reshape the issue of psychosocial factors identification accordingly, and thereby, strengthen the medical interest towards the findings of this paper.

References

- [1] R. G. Rodrigues, R. M. Das Dores, C. G., Camilo-Junior, T. C. Rosa, SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks, *Int. J. Med. Infor.* 2016;85(1):80-95.
- [2] L. Zhang, M. Hall, D. Bastola, Utilizing Twitter data for analysis of chemotherapy”, *Int. J. Med. Infor.* 2018;120:92-100.
- [3] S. Bao., S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, Y. Yu. Mining social emotions from affective text, *IEEE Trans. Knowl. Data Eng.* 2012;24(9):1658-70.
- [4] F. Ren, X. Kang, C. Quan, Examining accumulated emotional traits in suicide blogs with an emotion topic model, *IEEE J. Biom. Heal. Infor.* 2016;20(5):1384-1396.
- [5] A. Akay, A. Dragomir, B.E. Erlandsson, and M. Akay, Assessing antidepressants using intelligent data monitoring and mining of online fora, *IEEE J. Biom. Heal. Infor.* 2016;20(4):977-86.
- [6] B. Saha, T. Nguyen, D. Phung, and S. Venkatesh, A framework for classifying online mental health related communities with an interest in depression, *IEEE J. Biom.. Heal. Infor.* 2016;20(4): 1008-15.
- [7] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, M. Berk, Affective and content analysis of online depression communities, *IEEE Trans. Affe. Comp.* 2014;5(3):217-26.
- [8] C. H. Tai, Z. H. Tan and Y. S. Chang, Systematical approach for detecting the intention and intensity of feelings on social network, *IEEE J. Biom. Heal. Infor.* 2016;20(4):987-95.
- [9] S. Cohen, and S., L. Syme. Issues in the Study and Application of Social Support. In: S. Cohen and S.L. Syme, editors, *Social Support and Health*, New York: Academic Press, 1985. p. 3-22
- [10] Suomi24, Suomen Suurin Verkkoyhteisö, 2000, available at <https://www.suomi24.fi/> Accessed 2020
- [11] K. H. Lagus, M. Pantzar, M.S. Ruckenstein., M. J. Ylisiurua, Suomi24 -Muodonantoa aineistolle (Suomi24 - Giving shape to the data set). In: *Proceedings of the Valtiotieteellisen tiekunnan julkaisuja Workshop*, Helsinki; 2016.
- [12] C.R. Sheldon, R. Mermelstein, T. Kamarck, and H.R. Hoberman, Measuring the Functional Components of Social Support. In: I.G. Sarason and B.R. Sarason, editors, *Social Support: Theory, Research, and Applications*, Martinus Nijhoff. 1985. p.73-94.
- [13] J. Suler, The online disinhibition effect. *Cyber. Behav, Soc. Networks* 2014;7(3): 321–6.
- [14] P. E. Kummervold, D. Gammon, S. Bergvik, J.K. Johnsen, T. Hasvold, J.H. Rosenvinge, Social support in a wired world: use of online mental health forums in Norway. *Nord. J Psych.* 2002; 56(1):59-65.
- [15] A.J. Mitchell, M. Chan, H. Bhatti, M. Halton, L. Grassi, C. Johansen, N. Meader, Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: A meta-analysis of 94 interview-based studies. *The Lancet Onco.* 2011;12(2): 160–74.

- [16] K. M. Lagus, M. Pantzar, M. Ruckenstein, and M. Ylisiurua, Suomi24 – Muodonantoa Aineistolle. Technical Report, University of Helsinki, 2016.
- [17] F. Salvetti, and N. Nicolov, Weblog Classification for Fast Splog Filtering: A URL Language Model Segmentation Approach, In: Proceedings of HLT-NAACL Human Language Technology Conference, New York City, USA.2006.
- [18] E. Adar and L.A. Adamic, Tracking Information Epidemics in Blogspace. In: Proceedings of Web Intelligence, Compiegne, 2006.
- [19] M. Thelwall, Link Analysis: An Information Science Approach. Academic Press, 2004
- [20] J. Yi, Detecting buzz from time-sequenced document streams. In: Proceedings of the IEEE conference on eTechnology, e-Commerce and e-Service, (EEE '05); 2005, p.347- 352.
- [21] L. A. Adamic and E. Adar, Friends and neighbors on the Web. Social Networks, 2003;25(3):211-30
- [22] T. Lento, H. T. Welser, L. Gu, and M. Smith, The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System, In: Proceedings of the third annual workshop on the weblogging Ecosystem: Aggregation, Analysis, and Dynamics. Edinburgh, Scotland, 2006.
- [23] T. Kohonen, S. Kaski, K. Lagus, J. Salojäärvä, J. Honkela, V. Paatero, and A. Saarela, Self organization of a massive text document collection. Trans. on Neur. Networks. Special Issue on Neural Networks for Data Mining and Knowledge Discovery, 2000;11(3):574-85.
- [24] K. Lagus, M. Ruckenstein, A. Juvonen, and C. Rajani, Medicine Radar – A tool for exploring online health discussions. In: Proceedings of the Digital Humanities in the Nordic Countries, CEUR Workshop, Vol. 2084; 2018. p.460-468.
- [25] Suomi24. Suomi24 dataset corpus, available at <http://metashare.csc.fi/repository/download/b4db73da85ce11e4912c005056be118ea699d93902fa49d69b0f4d1e692dd5f1> Archived by WebCite at <http://www.webcitation.org/6oeEAXcAo>. Accessed June 2020
- [26] P. Lueboonthavatchai, Prevalence and psychosocial factors of anxiety and depression in breast cancer patients. J. Med. Assoc. of Thailand, 2017;90(10):2164.
- [27] B. Garssen, Psycho-oncology and cancer: linking psychosocial factors with cancer development. Ann. of Onco. 2002;13(4):171-5.
- [28] P. Reynolds and G.A. Kaplan, Social connections and risk for cancer: prospective evidence from the Alameda County Study. Behav. Med., 1990;16(3):101-10.
- [29] M.E. Peters, M.M. Goedendorp, S.A. Verhagen, W.T. Van-Der-Graaf, G. Bleijenberg, Exploring the contribution of psychosocial factors to fatigue in patients with advanced incurable cancer. Psycho-Oncology, 2014;23(7):773-9.
- [30] S. Moffatt, E. Noble, M. White, Addressing the Financial Consequences of Cancer: Qualitative Evaluation of a Welfare Rights Advice Service. PLoS ONE 7(8): 2017; e42979, <https://doi.org/10.1371/journal.pone.0042979>
- [31] V.S. Helgeson, S. Cohen, R. Schulz, J. Yasko, Education and peer discussion group interventions and adjustment to breast cancer. Arch. of Gen. Psych., 1990;56(4):340-7.
- [32] L. Northouse, A.L. Williams, B. Given, R. McCorkle, Psychosocial care for family caregivers of patients with cancer. J. of Clin. Oncol. 2012;30(11): 1227-34.

- [33] A. Schmidt and M. Wiegand, A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics, Valencia, Spain, 2017. p. 1–10
- [34] FinnWordNet, 2019, available at <http://www.ling.helsinki.fi/en/lt/research/finnwordnet/> Accessed 2019
- [35] Z. Wu, and M. Palmer, Verbs semantics and lexical selection, In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994. p.133-138.
- [36] A. L. Falisi et al., Social media for breast cancer survivors: a literature review. J Cancer Surviv, 2017;11: 808–21.
- [37] D. C. McFarland and J. C. Holland, The management of psychosocial issues in oncology. Clin. Adv. in Hemat. & Oncol. 2016;14:999–1009.
- [38] Mindscapes. Citizen Mindscapes Project. <https://blogs.helsinki.fi/citizenmindscapes/> , accessed 2019
- [39] M. Mohamed and M. Oussalah, SRL-ESA-TextSum. A Text Summarization Approach Based on Semantic Role Labeling and Explicit Semantic Analysis, Infor. Proc. and Manag. 2020, <https://doi.org/10.1016/j.ipm.2019.04.003>
- [40] M. Mohamed and M. Oussalah, A Hybrid Approach for Paraphrase Identification Based on Knowledge-enriched Semantic Heuristics, Lang. Res. and Eval. 2019, <https://doi.org/10.1007/s10579-019-09466-4>

Figures

Figure 1

Overall approach for data collection

Algorithm 1 repetition detection from Suomi24

Input finnishCancerKeywords, Suomi24 files,
rightTopics=[Health], finnishStopwords, FreqDist
from NLTK

Output graphs, mostRepeatedwords array,
cancerRepeated array, outputFile

```

1: procedure REPETITION DETECTION
2:   for every file in Suomi24 files do
3:     items ← file/items
4:     for every item in items do
5:       topics ← item/topics
6:       for topic in topics do
7:         if topic in rightTopics then
8:           body ← item/body
9:           if body contains finnishCancerKey-
words then
10:            Sentences ←
Divide sentences in TranslatedTweet;
11:            Tokens ← Tokenize words in Sen-
tences;
12:            totalDist ← FreqDist(words)
13:          end if
14:        end if
15:      end for
16:    end for
17:  end for
18:  for every item,value in totalDist do
19:    if item in finnishCancerKeywords then
20:      cancerRepeated ← [item,value]
21:    end if
22:  end for
23:  mostRepeatedwords ← totalDist[0:12]
24:  function VISUALIZE
25:    x-axis ← (mostRepeatedwords,cancerRepeated) items
26:    x-axis ← (mostRepeatedwords,cancerRepeated) values
27:    VISUALIZE
28:  end function
29: end procedure

```

Figure 2

Frequent Cancer term detection algorithm

Komission täytyy pyytää selvitystä ministeriltä ja hänen avustajaltaan .
Comission must ask_for clarification from_minister and his assistant .

Figure 3

Algorithm 2 co-occurrence detection and topic from Suomi24

Input: Finnish Cancer Keywords, Suomi24 files, Right Topics, Finnish Stopwords, topics Keywords, database Access Keys
Output: wordCount[topics,wordDistance,isComment,keywords], database

```

1: procedure SUOMI24_ANALYSIS
2:   for every file in Suomi24 files do
3:     items ← file/items
4:     for every item in items do
5:       topics ← item/topics
6:       body ← item/body
7:       for sentence in body and comments do
8:         for word in sentence do
9:           if length of word > 2 and length of the next
word > 2 then
10:            wordCount ← word cooccurrence
11:            if distance between words < 3 then
12:              wordCount[wordDistance] ← "1-2"
13:            else if distance between words < 6 then
14:              wordCount[wordDistance] ← "3-5"
15:            else if distance between words < 9 then
16:              wordCount[wordDistance] ← "6-8"
17:            else if distance between words < 15 then
18:              wordCount[wordDistance] ← "8-14"
19:            else
20:              wordCount[wordDistance] ← "15+"
21:            end if
22:            if word in deathList[] then
23:              increment keywords["death"]
24:            else if word in illnessList[] then
25:              increment keywords["illness"]
26:            else if word in treatmentList[] then
27:              increment keywords["treatment"]
28:            else if word in socialList[] then
29:              increment keywords["social"]
30:            else if word in financialList[] then
31:              increment keywords["financial"]
32:            end if
33:            wordCount[keywords] ← keywords
34:          end if
35:          if inComments then
36:            wordCount[isComment] ← inComments
37:          else if inBody then
38:            wordCount[isComment] ← inTopic
39:          end if
40:          wordCount[topics] ← topics
41:          DATABASE ← wordCount/topics,wordDistance,isComment,keywords]
42:        end for
43:      end for
44:    end for
45:  end for
46: end procedure

```

Figure 4

Co-occurrence detection and topic modeling

Figure 5

Histogram of most co-occurring terms and comparison with cancer-related terms with distance

Figure 6

Overall architecture of psychosocial category ontology generation

Count (%) vs. Keywords category

Figure 7

Psychosocial factor categorization percentage

Search Suomi24 database

Key words: Terveys

Instructions:

Key words
Keywords included in any word: Financial, social and treatment related.

Topics
Search co-occurrences for such topics (in Finnish).

Word distance
Search co-occurrences with such word distance, options: 1-2, 3-5, 6-8, 9-14, 15+

Search outcome for the query example

Item	Keywords
270. terveysviestinta2014-terveys	(social:59,treatment:59)
271. terveysviestinta2014-uta	(social:60,treatment:60)
272. terveysviranomaiset-omaga3	(death:1,financial:7,illness:1,social:12,treatment:4)
273. terveysviranomaiset-alku	(death:1,financial:7,illness:1,social:12,treatment:4)
274. terveysviranomaiset-esim	(death:1,financial:7,illness:1,social:12,treatment:4)
275. terveysviranomaiset-gramma	(death:1,financial:7,illness:1,social:12,treatment:4)
276. terveysviranomaiset-paivasa	(death:1,financial:7,illness:1,social:12,treatment:4)
277. terveysviranomaiset-rasvaha	(death:1,financial:7,illness:1,social:12,treatment:4)
278. terveysviranomaiset-suostele	(death:1,financial:7,illness:1,social:12,treatment:4)

Figure 8

Search interface from the web application. Left: Search Interface from web application. Right: Search outcome for the query example

Figure 9

Sample output on database (Firebase)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryfile.docx](#)