UNIVERSITY OF OULU

FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Moamen Ibrahim

Revealing effects of psychosocial factors of cancer patients

Master's Thesis
Degree Programme in Computer Science and Engineering
January 2019

Ibrahim M. (2019) Revealing effects of psychosocial factors of cancer patients. University of Oulu, Degree Programme in Computer Science and Engineering, 73 p.

ABSTRACT

This research shows different methodologies applied on different platforms in order to extract both social and psychosocial factors that might be related to caner by applying natural language processing tools on text from different platforms as social media or other online forums. We also present challenges associated with every platform and the corresponding tools used on it. From text mining to text analysis and then data visualisation, this research compares different analysis methods and outputs. We discuss many tools either tested, used or modified in order to achieve such analysis. Meanwhile, we were able to get interesting findings for the medical fields to explore and research more. We developed a modular system that can help clinicians and medical experts use to analyse similar forums.

Keywords: text mining, text analysis, psychosocial factors, cancer, natural language processing.

Ibrahim M. (2019) Syöpäpotilaiden psykososiaalisten tekijöiden vaikutusten paljastaminen. Oulun yliopisto, Tietotekniikan tutkinto-ohjelma, 73 s.

TIIVISTELMÄ

Tämä tutkimus esittelee erilaisia menetelmiä sovellettuina eri alustoilla, tavoitteena hahmottaa sekä sosiaalisia että psykokososiaalisia tekijöitä, jotka voivat liittyä syöpään sovellettaessa luonnollisia kielenkäsittelyvälineitä eri alustojen tekstille sosiaalisen median tai muiden online-foorumeiden muodossa. Esitämme myös haasteita, jotka liittyvät jokaiseen alustaan ja siihen liittyviin työkaluihin. Teksti-mining, tekstianalyysiin ja sitten datan visualisointiin tässä tutkimuksessa verrataan erilaisia analyysimenetelmiä ja -tuloksia. Keskustelemme monista työkaluista, jotka on testattu, käytetty tai muunnettu tällaisen analyysin saavuttamiseksi. Samaan aikaan saimme mielenkiintoisia tuloksia lääketieteen aloille tutkia ja tutkia lisää. Kehitimme modulaarisen järjestelmän, jonka avulla lääkärit ja lääketieteen asiantuntijat voivat analysoida samanlaisia foorumeita.

Avainsanat: tekstin mining, tekstianalyysi, psykososiaaliset tekijät, syöpä, luonnollinen kielenkäsittely.

TABLE OF CONTENTS

FOREWORD

I would like to thank prof. Mourad Oussalah for acting as the principal supervisor of this thesis and Ekaterina Gilman for their technical supervision. I would also like to thank Dan Ackers for his help and support when starting this research. Thank you Mohamed Aboeleinen, Mina Gabriel, Mostafa Abd Al-Kareem, Mohamed Khalifa, Hassan Mehmud, Sercan Turkmen, Katri Kurkela for your enthusiasm and general support. When I first started my masters degree, I had different ideas how my thesis might be. But I usually have this idea that I enjoy for a while and then it fades away. I needed to have an idea that will motivate me to keep going and keep track of the target of the study. I found that in this study, as when I usually start to feel less motivated, I remember how important this study can be for many people, not only researchers, but clinicians, patients and even those who lost someone to cancer, as my I did to my grandmother. I hope this research helps for better future for such field and I can not be very thankful to everyone who stood by my side to make it possible, from friends here and in my home country to my family.

Oulu, 19th June, 2019

Moamen Ibrahim

LIST OF ABBREVIATIONS AND SYMBOLS

CSE	Computer Science and Engineering
NLP	Natural language processing
NLTK	Natural language processing toolkit
NLU	Natural language understanding
LCSM	Lung Cancer Social Media
BCSM	Breast Cancer Social Media
CHAT	Cancer Hallmarks Analytic Tool
SEER	Surveillance, Epidemiology and End Results
RSS	Really Simple Syndication
NLTK	Natural language processing toolkit
EHR	Electronics health records
API	Application program interface
JSON	JavaScript Object Notation
CSV	Comma-separated values
REST	Representational state transfer
EULAs	End User License Agreements
SNA	Social network analysis

1. INTRODUCTION

1.1. Background and motivation

Many lives are affected by cancer worldwide, it kills more people in the United States than accidents, stroke, Alzheimer's and diabetes combined. Despite spending 147 billion dollars on cancer care in 2017, 600 thousand people died from cancer and 1.7 million new people were diagnosed with cancer as well. The great need for optimising the well-being and lifestyle of cancer patients is one of the main focuses of medical experts in the field. Improving the mortality rate is the direct result of advancement in cancer detection and cancer care. Clinical trials give researchers more insights as they always seek the improvement of the current treatment given to the patients. Such trials can offer a way to measure the success and failure of treatments, which can be beneficial for future research. What helps researchers in such states is that patients are usually engaging and motivated to their clinicians and their treatment decisions.

The challenge for such trials is to get the patients that would like to participate in such studies mainly because of concerns about side effects, costs, location and unproven treatment. Such challenge is getting deeper and harder as patients tend to social media to express there experiences and daily challenges in life these days. This development encourages for more focus on social media for such field. Physicians need to create a visible presence online to encourage patients to participate in such studies. Such attempts made good progress and with the rise of social media, now such platforms already caught the attention of researchers in such fields.

Online communities such as the LCSM (Lung Cancer Social Media) Twitter group and the BCSM (Breast Cancer Social Media). Twitter group both host weekly Twitter-based chats to educate patients about lung cancer. Such platforms were started by and maintained by physicians and patients and have had a strong impact on patient's education. Currently, many clinical trials are available only at large cancer hospitals because they have the resources to run, open, and maintain trials. While social media can give us more access and information to various aspects that can be challenging to conduct on smaller scales or not in a large cancer hospital.

With many users communicating as patients and clinicians these days on social media and other online platforms, it makes sense to utilise and generate tools that can extract information and provide it to decision makers and those interested in more analysis and research in such field. With this breakthrough happening to humanity in the field of social media these days, we seem to be heading towards and unknown territory of psychological and psychosocial dynamics in the human experiences in a fast pace. Research in such fields does not seem to cope up with the fast technology wave that influenced our society in different aspects and directions. Great advantage of this is that internet and social media have allowed connections and free flow of ideas that were never possible to this vast extend before. Also, people become less shy and can connect with other they normally would not meet, but there is also disadvantages to such tools. It can sometimes feel shadowy there where people can hide behind anonymity and not have that real social connections with others.

Researchers are continuing on examining the psychological fallout and tendencies of the rapidly growing internet and social media universe, and provide caution as needed to the public. One of such aspects is the pyschosocial support from family members, relatives and friends to cancer patients that are now using social media rapidly with the rise of social media use. This inspired us to tackle a specific target in such vast track.

1.2. Research questions and scope of thesis

This study examines the psychosocial factors extracted from social media and online platforms. We managed to compile a list of research questions where we will focus in this research to tackle.

The first question is: What is the variability in which people in Nordic region use twitter to communicate in cancer related matters? It was addressed by the amount of tweets fetched in the Nordic region that contains text related to cancer. This is handled in fetching and filtering tweets subsections in the data collection part for streaming tweets approach. In the data collection part, we narrowed down our dataset range to include tweets from the Nordic area only. Also, we took a chance to analyse a specific user and check for variations and changes.

The second question is: How to identify psychosocial factors from the Twitter dataset? The question can be answered using natural language tools and approaches that can enable us to identify factors using matching of lexical findings with bag of words and using also modern tools as we identified in the data analysis part. Where we try to identify the key parameters that affects psychosocial factors as we try to extract key findings as topics, sentiment, most used words, hashtags, part of speech tags and named entities. It is challenging to identify psychosocial factors from unstructured text as it is on Twitter, so we had to use combination of tools and construct our own methedologies to achieve such requirements.

The third question is: How to track the evolution of the disease and outcomes through Twitter dataset? Trying to find an answer for the third question was hard as most of the times users were not open about the status of there disease. But we did some analysis on the overall of user profile, who was diagnosed and survived from cancer in this research.

The fourth question is: How to track psychosocial factors from online forums related to cancer and compare those to Twitter? Comparing the findings from the online forums to those of Twitter can give a proper understanding about the variations of how users communicate about cancer related issues on such platforms. The comparison should include the focus of the study, also, which is the pyschosocial factors related to these, including the topic and sentiment analysis.

The fifth question is: How to track the evolution of the disease and outcomes through online forums related to cancer? This is done by identifying list of users and fetching their posts and threads for all the period of there profile, trying to identify variations and differences in their posts.

1.3. Research methodology

In this research, we focus on the comparison between two different social media analysis and online forums analysis. In the first approach, we applied natural language processing tools on tweets from different places but restricted to the Nordic area, while for the other one we focused on a specific user. We wanted to discover the challenges, constrains and also advantages of each approach. While for the last approach, we wanted to apply what we learned already from the other two approaches and apply these to the new one. It will prove different findings and also interesting results. While also comparing the performance of every tool used and comparing its output in comparison to different methodologies.

1.4. Scientific contribution

The research suggests different approaches to extract pyschosocial factors from cancer patients online. Comparing every approach with the performance and also the findings will help us understand which approach is the best for every application. The research paves the way for future researchers in such new approach in research. It is challenging to fetch such information from a free text online as social media and free discussion threads. These obstacles had to be investigated to provide future guidance for research in such field.

1.5. Structure of thesis

Thesis is divided into seven sections, where in the first section, we introduce the research motivation, background and main methodology. While in the second section, we introduce the state of the art of such research and how different research approaches were not intended for such application. Main topics were analysed as the main psychosocial factors and their associations, psychosocial variables and how to measure them, cancer oncology, some reflections in specific cancer types and tools developed before in such fields. In was also important to highlight the Nordic area social media activity and how to tackle that in our research. Natural language processing is also our focus where we use different tools and techniques to achieve such research. These tools were discussed in detail in section three in the thesis. Section four explains our implementation as it is divided into three parts as explained previously; Nordic area tweets analysis, specific Twitter user analysis and online forum discussion analysis. Also, our implementation of an online platform to show our output to proper users. Section five show case our results in every approach with both interesting findings and challenges discussed. In section six we expand more on these results and show bottlenecks, possible mistakes and suggested approaches to tackle them, while in section seven we conclude these findings from the research.

2. BACKGROUND AND RELATED WORK

2.1. Psychosocial factors associations

Psychosocial factors have previously been linked with survival and mortality in cancer populations. Nevertheless, little is known about how such relationship occurs and develops as compared to the well studied biological factors. Therefore, models have been developed that outline the potential mechanisms by which psychosocial factors may be associated with clinical outcomes in cancer.

Primarily derived from empirical research, [1] posits that psychosocial factors, such as stress and coping, are associated with timorousness through the effects of stress hormones on immunity. This was motivated by findings that higher levels of perceived stress and chronic stress have been associated with greater likelihood of developing pre-cancers [2], shorter disease-free interval following cancer treatment [3] and increased risk of cancer recurrence [4].

Similarly, high levels of social well-being and social support have been linked with lower levels of vascular endothelial growth factor and proteinase factors known to stimulate tumour growth and progression [5]. [6] have shown that greater depressive symptoms and depressive or avoidant coping styles are associated with increased mortality. Faller & Bulzebruck [7] provided empirical evidence that active coping styles, emotional support [8] and better global quality of life are associated with a decreased risk of mortality and longer survival across several cancer populations.

However, the above findings are contrasted by other research that refutes the positive relationship between psychosocial factors and mortality across cancer populations, see for instance, [9], especially in case of stressful life events. Despite the debate on the effect of psychological factors on some specific cancer population, there is evidence that suggests that implementing psychotherapeutic interventions to reduce distress is associated with reduced risk of developing pre-cancer [10] as well as longer survival times and longer time to disease recurrence in breast cancer [11]. These findings dervies the need to identify the actual psychosocial variables and how to measure it.

2.2. Psychosocial Variables and Measures

Defining the psychosocial variables and measures is very crucial for such study, the categories of psychosocial factors and instruments used to access them should be defined before conducting such study. [12] Explains how the variables were stated, depression and anxiety based on a pre-study using questionnaires. The measures used to assess childhood environment (e.g., the quality of relationships with family members, the extent of household responsibilities during childhood). Family Relations and Childhood Memories Questionnaire, maternal closeness ratings based on responses to the Thematic Apperception Test [13].

Personality style is also another aspect, conflict avoidance can be one example, where patients tend to forget about their state or find a way to escape it, making it a factor to affect the state of the patient, same for denial and repression [12]. Expression of anger and hostility can affect psychology and sociology of the cancer patient, where it can be shown in stressful life events and separation or loss [13]. Measures of extroversion and introversion were discussed also to have significant effect on psychosocial matters [13]. It will be interesting to discover cancer oncology itself and how it is related somehow to our research.

2.3. Cancer oncology

Cancer immunotherapy has reached an important level in curing cancer, different therapy means have reported a consistent response for broad range of human cancers with different agents [14]. Cancer is characterised by the accumulation of multiple genetic alterations and the loss of regularly tissue and cell development, this urges the need to examine the cancer immunity cycle carefully [14]. The goal of cancer immunotherapy is to initiate and re-initiate a self-sustaining cycle of cancer immunity, enabling it to amplify and propagate, they must be carefully configured to overcome the negative feedback mechanisms. Medical experts who help in diagnosis of cancer, staging the cancer and grading the aggressive nature of the cancer are called oncologists who usually use tools like the medical history of the patient. Oncology depends on tools like biopsy which is the removal of bits of the tumour tissue and examining it under the microscope, endoscopy for the gastrointestinal tract, imaging studies like X-rays, CT scanning, MRI scanning, ultrasound and other radiological techniques, Scintigraphy, Single Photon Emission Computed Tomography, Positron emission tomography and nuclear medicine techniques [15].

The effect of such treatment on the cancer patient have many complications, one of these affects is called cancer fatigue which is considered as a persistent subjective sense of tiredness related to cancer treatment that affects the usual functioning. Studies suggest that it is amongst the most common symptoms experienced by cancer patients who are affected by radiotherapy and chemotherapy, among patients treated with chemotherapy or radiotherapy, more than one third mentioned that fatigue affected their ability to work, relationships with others, and physical and emotional well-being [16]. Studies indicate that psychological intervention from professionals, relatives and friends might affect or alleviate the effects from these kinds of therapy which can be an important factor for such patients to go through the different and hard phases of such disease. It is important as part of this study to discover the effects of such treatment on the cancer patients as these changes in mood, emotions and well-being should affect the way they behave in their psychological and social life. Specific cancer type is also important to highlight, as one of the most common cancer worldwide, it is expected that researchers will investigate breast cancer heavily.

2.4. Reflections in breast cancer

A meta-analysis examined the relationship between psychosocial factors and the development of breast cancer where strongest support was found for the hypotheses that breast cancer patients use a coping strategy based on denial in response to life stressors, have experienced separation and loss, and have a history of stressful life experiences [12]. The fact that breast cancer can take several years to develop and be detectable makes studies examining recent life stresses might be expected to yield non-significant results [17].

Although these studies provide support for the relationship between psychosocial factors and mortality in cancer, it does require specific clinical setting and procedure in order to interview clinical patients and their relatives. It may raise further challenge and psychological barrier.

2.5. Cancer research significant software tools

Currently, to the best of our knowledge, the best attempt at a similar concept was created in the University Hospital of North Norway (UNN), in 2016 [18]. They attempted to create patient trajectories to determine the progression of diseases and disease types, using these trajectories, they were able to detect 80% of the patient events ahead of time, all using only free text. They believe that their method can be used as data driven decision support tool that can be used during the complete cancer patient trajectory. There are various other researches that have been carried out in topics surrounding the data mining of cancer in attempts to create models around cancer related topics [18].

A paper discussed the research literature on text mining to find cancer domains, the research suggested that it is very important to use more machine learning models instead of rule-based methods. It also stated the challenge of small training data-sets in the clinical domain [19]. As illustrated in Cancer Hallmarks Analytic Tool (CHAT) developed to to organise and evaluate scientific literature on cancer. CHAT is capable of retrieving and organising millions of cancer-related references from PubMed into the taxonomy [20]. Their natural language processing pipeline had a structure that can be a good starting framework for our analysis in extracting features and understanding out of cancer related text [20].

Surveillance, Epidemiology and End Results (SEER) program of the National Cancer Institute collects data on cancer diagnoses, treatment and survival for approximately 30% of the United States (U.S.) population [21]. The main point of this program is to reflect from these huge amount of data on the cancer and oncology research and practices in medical centres and also collecting significant amount of data on the pathology diagnoses across demographic groups, geographic regions, and time, and providing unique insights into the practice of oncology in the U.S that are not attainable from other sources

[21]. The program provides good data for from a huge resources to analyse the incidence, survival and mortality data for histopathologic cancer sub-types. The program seems successful in developing next generation tools for analysis and cancer research improvement [21]. Modern approaches to handle cancer are emerging, one of the emerging technologies is "The Cancer Genome Atlas" merging molecular data with histopathological diagnosis meaningful cancer classifications is a central goal in cancer control and is redefining the practice of oncology [21].

Data collected for all primary invasive cancers and some other diagnoses include date of diagnosis, and demographic variables such as age at diagnosis, gender, race/ethnicity, and county of residence. This made the program to have good variety of data for different analysis and further studies [21]. The SEER program provides huge information for frequency distributions, incidence and mortality rate over time on all cancers. It provides data suitable for comparative analysis of cases within populations of defined characteristics [21].

Table 1. Tools developed in cancer research

Author	**Tool name if available**	**Methodology**	**Result**	**Findings**
University Hospital of North Norway (UNN)	-	Analysis of free text of around 4,080 cancer patients, methodology allows disease trajectories of the cancer patients to be estimated from free text in electronic health records (EHRs)	predict 80% of patient events ahead in time	Helpful tool as it improves clinical decision support and personalise trajectories, thereby decreasing adverse events and epitomising cancer treatment
University of Cambridge and Karolinska Institutet	CHAT	Using an NLP pipeline to develop a text mining tool capable of retrieving and organising millions of cancer-related references from PubMed into the taxonomy	It offers a great potential to organise and correctly classify cancer-related literature	Useful tool in identifying hallmarks associated with extrinsic factors, biomarkers and therapeutics targets.
National Cancer Institute	SEER	Collect amount of data from different demographics, groups, geographic regions and time to get specific significance	Successfully build huge distributions, incidence and mortality rate over time on all cancer types	The huge amount of data available helped the project progress to lots of findings and analysis

2.6. Effects of social media platforms

In the era of social media platforms, including Twitter, Facebook, blogs, which invaded all population groups and fields. Health researchers are provided by golden opportunities to access to patients critical thoughts, feelings, experiences and worries, without any ethical nor physical barrier. This emerging form of unsolicited communication among self-forming online communities of patients and individuals with diverse health concerns is referred to as peer-to-peer support [22].

Strictly speaking, Twitter messages and network communities have been investigated in many related studies [23]. In micro blogging services such as Twitter, users may be overwhelmed by raw data, Some researchers suggested to solve this problem by classifying short messages which can give access to sufficient word occurrences and methods that have limitations such as "Bag-of-Word". The study proposed a system where users can subscribe to specific types of tweets and messages [24]. Some researches used the real-time nature of Twitter to detect events using machine learning models and send messages to those interested to receive such information [25].

Intuitively, cancer patient Twitter users share treatment experience, clinical effectiveness, financial burden, family worries, lifestyle with other patients, close friends and relatives. In America, users are becoming more open in using Twitter to share their experience with cancer, which gives chance for older generations and younger ones who feels less anxiety using computer-mediated communication to share their views and experience. Some cancer patients share their daily moods and feelings.

For example, "Lisa Bonchek Adams", a breast cancer shared her experience with over 176,000 tweets. In some of her tweets, she was explaining what the chemotherapy and radiation is doing to her, she died in April 2015, while her tweets left a great help for analysts [26]. As length of tweets is relatively short compared to blogs and other popular media for cancer patients including hospital materials, patients are more likely to tweet about their personal struggle on Twitter [27]. Besides, many health authorities have also started using social media platforms, including, twitter, to communicate directly to patients and interact with them [28]. In one review conducted by Stanford, they considered the use and potential of social media and mHealth technologies for cancer prevention, cancer treatment, and survivor-ship, with clear advantages in broad reach, scaled delivery and low resource setting, health authorities can develop supportive social networks, connect patients and providers, encourage adherence with cancer care, and collect vast quantities of data for advancing cancer research [28].

Cancer, however, is one of the most sensitive issues that patients generally do not feel comfortable writing about and hence it is a challenge for researchers to detect them and perform analysis to find potentially important metrics for clinicians, doctors and nurses or someone else that could be able to help. This opens up interesting challenges for information processing community to identify relevant cues that may lead to better identification and comprehension of psychosocial factors.

Table 2. Different analysis on Twitter

Author	**Methodology**	**Findings**
Ohio state university	Overcome raw data challenge by classifying short messages which can give access to sufficient word occurrences and methods that have limitations such as “Bag-of-Word"	BOW approach performs decently but 8F performs significantly better with this set of generic classes.
University of Tokyo	classifier of tweets based on features such as the keywords in a tweet, the number of words, and their context, each twitter user is a sensor and then applying Kalman filtering and particle filtering	The particle filter works better than other compared methods in estimating the centres of earthquakes and the trajectories of typhoons

2.7. Social media in Nordic area

This motivates the current work, which tries to shed the light on the psychosocial factors of Nordic cancer population through Twitter analysis. According to previous research about social networks in Nordic region, the Nordic network holds a solid and cohesive structure proved by their diameter and inter-distances. It is set up by three well-defined sub-networks: the Finnish, the Sweden and the Danish sub-network [29].

Finnish network is characterised by the strong presence of the University of Helsinki, which is the most linked Finnish university and the authentic gateway between Finland and Europe. The rest of Finnish universities have a quite low proportion of links and their come/go mainly from/to Sweden and United Kingdom [29]. Table 3 shows some statistics on social media platforms.

Table 3. Social media distribution in Nordic region

Social media platform	Total percentage of usage	Male percentage	Female percentage
Facebook	78%	72%	83%
Youtube	72%	77%	68%
Instagram	39%	31%	47%
LinkedIn	27%	31%	23%
WhatsApp	27%	25%	29%
SnapChat	26%	23%	29%
Twitter	19%	22%	16%
Pintrest	14%	7%	21%

2.8. Medical taxonomy in social media

Some research was conducted on cancer taxonomy in social media, while the definition of social media was defined by the centres for Disease Control and Prevention as social networking sites, blogs, microblogs, RSS (Really Simple Syndication) feeds, and online forums. The type of information shared through the social media related to cancer may differ than that shared for general health promotion in other health context [30]. Earlier studies on social media were conducted on various resources as Facebook for social networking and Youtube for video sharing, researchers also checked online blogs and their content where most of the discussions were related to cancer survivorship and treatment [30]. Virtual world as a second life was also discussed in one study to show how an interactive world can help to train physicians to share bad medical news, such as diagnosis of cancer. Studies mostly focused on breast cancer, prostate cancer and other cancer types, while most of the articles in breast cancer focused on the content analysis of online forums for emotional support and self-expression [30].

Mostly in cancer related articles, users discuss about cancer treatments (e.g., prescriptions, chemotherapy, radiation) as well as disease outcomes and expectations, also their concerns related to sexual distress, anxiety, and depression arising from their diagnosis and treatment [30]. One study found it challenging to extract medical taxonomy from social media but using a good methodology of a systematic review of articles published through October 2013 as they developed a comprehensive search strategy for 3 medical health care databases (PubMed, Web of Knowledge, CINAHL) and Google Scholar, with proper data collection and filtering. It discovered that most of the users used these platforms for expressing emotions, raise awareness about cancer, provide support for cancer survivors and caregivers, promote information sharing and problem solving, treatment discussion and also raising funds [30].

It is obvious that the evolution of the use of social media should benefit in a way or another those responsible for cancer care, the emerging research from such area suggested and highlighted the importance of research for improving behaviours and promoting well-being and quality of life like increasing cancer screening, providing support during chemotherapy and reducing fatigue, as well as making social media an appealing place to share thoughts and worries about health with accessible, engaging, and interactive for increasingly diverse audiences [30].

In this chapter, we highlighted the psychosocial factors affecting cancer and their associations. Such variables and how to accurately measure them while highlighting cancer oncology. Also, complications of cancer treatment on the patient and studies how to tackle these. Meanwhile, many tools were developed to handle such studies, from prediction models to analysis models to identify cancer factors and relations. But with the existence of many free text on the social media platforms, it makes it perfect place to use text mining, analysis and visualise these analysis results. Natural language processing tools already provided good approaches to handle such text. In the next chapter we will focus on these tools and how to harness them in cancer research.

3. NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) is a track of artificial intelligence and linguistics made to teach the machine human language and make it understand the words written or spoken, it is usually classified into Natural language understanding (NLU) and Natural language generation (NLG) [31]. NLP is considered as an overlapping field of computer science, artificial intelligence and computational linguistics which involves the interaction between computers and human languages. It is an ability of computer system including natural language analysis, understanding and generation concerned with all linguistic forms, activities, or methods of communication, such as dialogue, correspondence, reading, written composition, publishing, translation and verbal reading [31].

3.1. Natural language understanding

Machines receive text in a natural language form, then NLU comes and tries to comprehend what the text means. By analysing and understanding the nature and structure of each word inside the text, NLU is responsible of resolving several ambiguities present in natural language such as [31]:

- Lexical ambiguity: Words have multiple meanings.
- Syntactic ambiguity: Sentence having multiple parse trees.
- Semantic ambiguity: Sentence having multiple meanings.
- Anaphoric ambiguity: Phrase or word which is previously mentioned but has a different meaning.

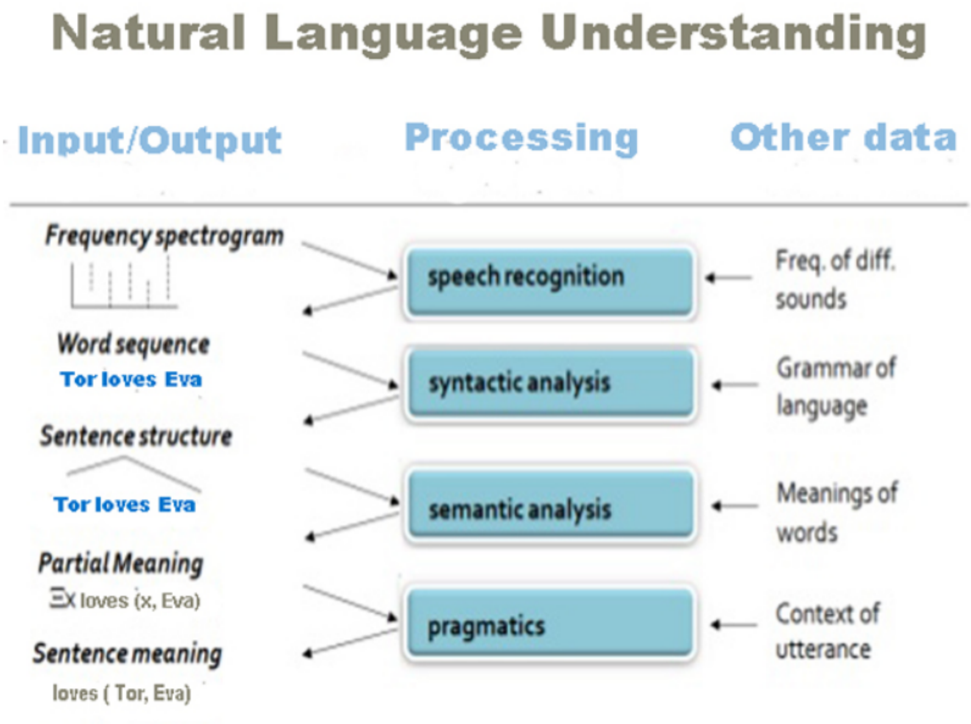


Figure 1. NLU Model (Image from www.maria-johnsen.com)

After that, the meaning of each word will be understood using lexicons and set of grammatical rules [32]. Maria Johnsen argued that [32], almost all recognition software is based on the Hidden Markov Models. They are models that is able to infer what you said through mathematical calculations shown in figure 1.

3.2. Natural language generation

NLG is the process of generating text from structured databases into a readable human language with meaningful phrases and sentences. With the help of the text-to-speech software, the system can create complete text. According to the [32], NLG can be divided into three proposed stages: text planning, sentence planning, realisation.

3.3. Process of NLP

The process of NLP mainly include five mains steps. Firstly, raw data is given in a text form where it should be broken into sentences in a segmentation step. The most important thing for this step is to know where to break. Second step is to break these sentences into words and punctuation, where this step is called tokenization. Natural language processing toolkit (NLTK) is an advanced platform for processing natural language in Python. The fourth step is entity detection, which is to chunk words, this needs to be done with tags or trees, so it is based on former results. And the last one is detecting relationships between entities

3.4. Tools for NLP

There are many tools used to analyse text, including tools for segmentation, tokenization, part of speech tagging, entity detection, relation detection, sentiment analysis and personality recognition.

3.4.1. NLTK

These tools are found in many programming languages, the famous tool is in Python which in Natural language Toolkit (NLTK), which is free, used widely and provides various free features. NLTK is written in Python and distributed under the GPL open source license. a broad-coverage natural language toolkit that provides a simple, extensible, uniform framework for assignments, demonstrations and projects. It is thoroughly documented, easy to learn, and simple to use. Over the past three years, NLTK has become popular in teaching and research [33].

3.4.2. Stanford tools

Stanford university have been developing many tools in the field of natural language processing. It provides a natural language parser as a program that works out the grammatical structure of sentences, for instance, which groups of words go together as phrases and which words are the subject or object of a verb. It is an extensible annotation-based NLP pipeline that provides core natural language analysis. This toolkit is quite widely used, both in the research NLP community and also among commercial and government users of open source NLP technology. It provides named entity detection and part of speech tagging as two of the important features provided by Stanford tools which are used extensively. Probabilistic parsers use knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences. These statistical parsers still make some mistakes, but commonly work rather well. Their development was one of the biggest breakthroughs in natural language processing in the 1990s [34].

3.4.3. Polyglot

One of the many challenges to face when dealing with non English text is checking for support for these languages such as Finnish. Polyglot provides support to many languages depending on the feature required [35]. Polyglot has many features including tokenization, language detection, named entity recognition, part of speech tagging, sentiment analysis, word embeddings, Morphological analysis and Transliteration [35]. Polyglot can be used on Finnish text to get named entities, which can detect locations, organisation and persons entities and supports 40 major languages.

3.4.4. Mallet

Mallet, a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text based on machine learning models which makes it smarter in analysing text [36]. Mallet includes sophisticated tools for document classification: efficient routines for converting text to "features", a wide variety of algorithms (including Naïve Bayes, Maximum Entropy, and Decision Trees), and code for evaluating classifier performance using several commonly used metrics.

3.4.5. Gensim

Gensim, as well, is considered a good tool for topic modeling for human speech and it is one of the most robust, efficient and hassle-free software that can categorise topics [37]. Gensim [37], however s a free open source code for text analysis, it is main target is to model topics out of free text, it follows an unsupervised semantic modelling from plain text. Latent Dirichlet allocation (LDA) is used to extract topics from texts using predetermined training material, provides non-destructive tokenization, named-entity recognition, part of speech tagging, string-to-hash mapping and all of this in a deep learning approach in Python. It is a matter of questioning how we are going to use this model to extract topics and what will affect on our process after extracting topics, taking in consideration that text is in different languages.

3.4.6. Gate

Gate [38] is an open-source tool in Java that provides many features and applications, developed in Java makes it a good tool to integrate with Hadoop to develop big data applications. Gate also introduced a graphical user interface which can be helpful for learning, research and testing. Such tool can give students, for example, hands-on experience towards natural language processing while doing constructive work in an enjoyable enviroment.

3.4.7. Turku parsers

A neural parsing pipeline for segmentation, morphological tagging, dependency parsing and lemmatization with pre-trained models for more than 50 languages, using Parser v2 and universal-lemmatizer (which uses Neural model for lemmatization using OpenNMT and pytorch libraries). Turku dep parser take over, the new pipeline is fully neural, it provides better accuracy, fast when parsing large documents, with about 5x faster than the previous Finnish-dep-parser as it takes advantage of GPU, the only downside for it is the long start-up cost when it's loading the models [39].

3.5. Cancer with NLP

Natural language processing accessed many fields and studies with wide range of appearance on the internet, cancer is one of these main topics that can be searched from the internet [40]. Searching for information on the internet is one of the main tasks performed by internet users searching on popular search engines and also on health information sites, by analysing 3 months of cancer-related queries on the Ask.com Web site, a prominent United States consumer search engine, which receives over 35 million queries per week, Overall 78.37% of sampled Cancer queries asked about 14 specific cancer types. Within each cancer type, queries were sorted into appropriate subcategories including at least the following: General Information, Symptoms, Diagnosis and Testing, Treatment, Statistics, Definition, and Cause/Risk/Link. The most-common specific cancer types mentioned in queries were Digestive/Gastrointestinal/Bowel (15.0%), Breast (11.7%), Skin (11.3%), and Genitourinary (10.5%), this indicates that natural language searching allows users to have the opportunity to fully express the information while it is a rising and expanding field [40].

The widespread use of electronics health records (EHRs) created rich databases for documenting the cancer patient's state with data such as cancer patient's continuum, while being locked in free text in an unstructured manner. The research developed at Brandeis University in Massachusetts concluded that using natural language processing on these records offers a promising method for structuring a free-text oncology history into a compact treatment summary, creating accurate and robust means of communication between patients and care providers [41].

Another research also tackled a different cancer identification through EHRs, where the research focused on identifying patients with prior colorectal cancer (CRC) testing, which is difficult to identify in patients. The research used NLP system to identify 4 CRC tests (colonoscopy, flexible sigmoidoscopy, fecal occult blood testing, and double contrast barium enema) within electronic clinical documentation, it was found that applying NLP to EHR records detected more CRC tests than either manual chart review or billing records. NLP had better precision to identify patients who were due for CRC screening than billing record review [42].

Researchers have always been interested in using natural language processing in exploring cancer in different platforms, from social media to EHRs and medical records. Also, exploring free-text in general which is a big advantage and a challenge for NLP at the same time. This study focuses more on using NLP in cancer in the social media domain, where free-text is abundant and also complex to define such topics, the study not only rely on identifying these cancer posts and threads, but also on identifying main issues concerning cancer patients in particular which can relate to the psychological and sociological aspects surrounding such patients.

4. IMPLEMENTATION

The analysis achieved to answer our research questions were made on two different online sources, Twitter as one of the leading social media platforms for any type of topic and also on online platforms specifically for cancer discussions. In the next section, we will discuss in details the technical implementation of both Twitter and online forum analysis. All implementation codes can be found online through Github[1].

4.1. Twitter analysis on Nordic area

Our focus here was to answer questions related to the Nordic region including Finland, Sweden, Norway, Denmark and Iceland. The main steps taken in both approaches are in a pipeline scheme. Starting from fetching tweets till sentiment analysis and natural language understanding. In the coming steps, we will explain the steps taken in twitter analysis. In the streaming tweets approach, we wanted to recognise the mostly used language in the Nordic region for such tweets about cancer. Also, when translating these tweets to English, we wanted to verify and make sure that our translation is valid and correct as well.

Twitter pipeline for streamed tweets, illustrated in figure 1, starts by fetching the tweets using stream listeners that collected tweets from the third of March till the 29th of it, the month of June, July and August as well. Using the stream listeners did not give us the option to limit these tweets to cancer related keywords, so we made some scripts to limit these tweets to cancer related keywords. The typical pipeline continues normally if the tweets were translated to English, if the tweets are in Finnish, some Finnish parsers and text analysis tools are used instead.

A research published by oxford university has proven that in order to make a research on twitter data sets it is important to measure these in different periods of time as users don't usually write about the specific topic in a specific period, so we might end up with a small data set of tweets including detectable cancer information [43]. As for our approach, we did our analysis first on tweets made in the Nordic region in March, we ended up with 3000 tweets maximum, but when adding data sets from June and July, we ended up with approximately 17200 tweets, with a big improvement in the tweets captured and improvement in our research material till 30000 tweets with August too.

The streaming approach starts by running scripts written in Python that acts as listeners to the Nordic region, whenever there is a new tweet detected in this region, the scripts catch it and save all information in a JSON format, as Tweepy library provides us these JSON formats. Then saved locally on our machines, we can then analyse each tweet and detect any cancer related word by matching

[1] Github links: `https://github.com/moamenibrahim/cancer-psychosocial-project` `https://github.com/moamenibrahim/discussions-text-analysis`

it with a predefined cancer keywords list, also extracting hash-tags, links and mentions. If the tweet language is Finnish, we run the Finnish tools, we improving on the other side of the project, which is originally developed by university of Turku, Finland and we are doing some improvements to it. Then we translate the tweet to English, which give us access to many tools available in English for natural language processing. Part of speech tagging, named entity recognition, hyponyms extraction, topic detection and sentiment analysis are detected using natural language processing tools integrated through our pipeline system, to be then dumped in JSON format with the extracted features and information and then uploaded to Firebase database that will then be our back-end database for a website, so that processed data can be viewed and accessed remotely. Also, categorisation, cancer type and stage detection are independent scripts that can be runned without interfering with other natural language processing tools used. At the end using other scripts, we were able to parse JSON files and plot the results that are displayed in this document further. As shown in Algorithm 1, Twitter pipeline with all processing of tweets and operations performed.

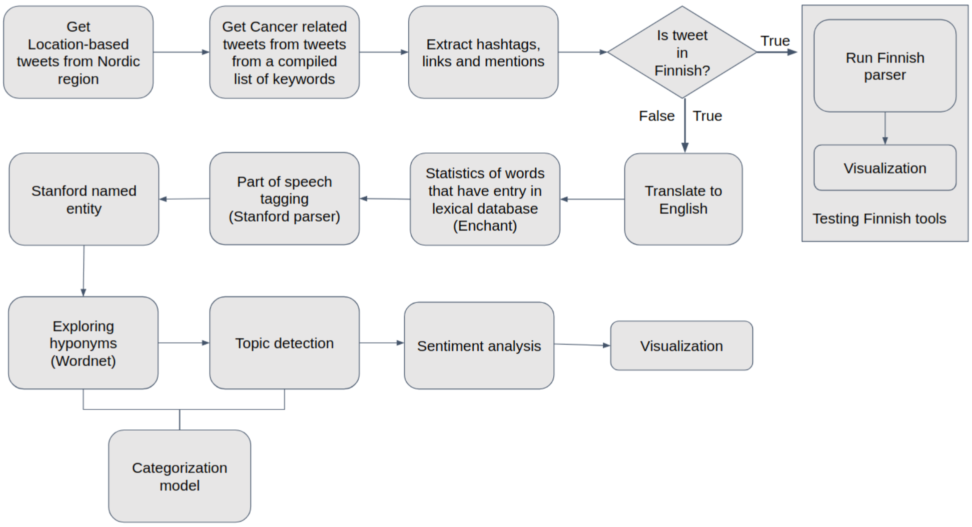

Get Location-based tweets from Nordic region
Get Cancer related tweets from tweets from a compiled list of keywords
Extract hashtags, links and mentions
Is tweet in Finnish?
True
Run Finnish parser
Visualization
Testing Finnish tools
False
True
Translate to English
Statistics of words that have entry in lexical database (Enchant)
Part of speech tagging (Stanford parser)
Stanford named entity
Exploring hyponyms (Wordnet)
Topic detection
Sentiment analysis
Visualization
Categorization model

Figure 2. System pipeline architecture

4.1.1. Data collection

Catching tweets using twitter streaming (Tweepy), which is an open-source python library, to catch tweets in Nordic countries in different languages, this step gave us insights on how much people use twitter in general which indicates that it is not as much as most of the other places in the world. The streamed tweets can reach by maximum 20k tweets per day in total, in the united states the average is 1 million tweets per day and more than 100k in Turkey, which can give us an indication from the beginning that twitter is not used that much in the Nordic region. Keeping in consideration that these numbers are limited to the number of calls given by the Twitter API, which forced us to make our

Algorithm 1 twitter pipeline

Input list of cancer related keywords k[], APIs authorisation tokens, location for streaming tweets longitude and latitude, Directory to save tweets outputDir.

Output (Topic, NamedEntity, Categorization, PartOfSpeechTags, Dictionary, Sentiment, TweetLength) Histograms, DataFrames and JsonFormattedFiles for results pushed to Database.

```
procedure PIPELINE
    while longitude and latitude in Nordic Region do Save Tweets in JsonFormattedFiles in outputDir;
    end while
    for every file in outputDir do
        for every Tweet in file do
            Extract and remove Hashtags, links and mentions;
            if any of k[] in Tweet or Hashtags then
                Get language from Tweet;
                if language = Finnish then
                    Run Finnish-deparser;
                end if
                TranslatedTweet ←translate Tweet to English;
                for word in TranslatedTweet do
                    if word in dictionary then
                        INCREMENT HitsInDictionary;
                    end if
                end for
                Sentences ← Divide sentences in TranslatedTweet;
                Tokens ←Tokenize words in Sentences;
                for every Token in Tokens do
                    PosArray[] ←Get part of speech tags;
                    NERArray[] ←Get named entity recognition;
                    Hyponyms[] ←Get Hyponyms for every word;
                end for
                TranslatedTweet ←Remove stopwords, punctuation from TranslatedTweet;
                Topic[] ←Generate topics from TranslatedTweet using LDA;
                Sent[] ←Evaluate sentiment from TranslatedTweet using SentiStrength;
            end if
        end for
    end for
    Dataframes, JsonFormattedFiles ←TranslatedTweet, Topic[], Sent[], Hashtags, Links, Mentions, Tweet, HitsInDictionary, PosArray[], NERArray[], Hyponyms[];
    Visualize DataFrames to histograms;
    Upload JsonFormattedFiles to Database;
end procedure
```

scripts dynamic and fault tolerant to some extent so they can endure longer to fetch bigger number of tweets using Tweepy, we are allowed to do up to 350 requests per hour to the Twitter API.

Using the same keywords used in the previous approaches, we conducted some manual searches using Twitter search API to get more understanding about the statistics and average numbers. In the following picture is the output of scripts using twitter search API `twitter.com/search`, which is a good tool that can provide till 7 days of historical data of tweets. The search API was able to catch tweets from different regions not only from the Nordic region, with a wider range in language used. Adding English, which improved some cancer types identification. Bone takes the lead in the Search API, while stomach is in the streaming one. Surprisingly, bone is the least in the Streaming API. While head and neck cancer was the least in search API approach, it did not have the same outcome in the streaming approach.

4.1.2. Formatting and cleaning data

The next step is to filter tweets depending on a compiled list of cancer related keywords, the list contains general keywords about cancer in English and other Nordic region languages, the used translation API is then validated by checking if these words have entries in the dictionary or not. The resultant number of tweets in the specified period, 3rd of March till the 29th was nearly 3000 tweets and by adding more data set from June and July, it increased to 17200 tweets and to 30000 by adding August tweets. Taking advantage of this, we were able to get the part of speech tagging and entity detection developed by such tools as shown in the following sample results. Figure 3 shows that using our model of keywords matching, we were able to identify that 1.1% of the total tweets in the Nordic region were related to cancer topics, out of the huge data set of streaming tweets, nearly 1.6 million tweets in total.

```
{
  "tweet": 911,
  "lang": "da",
  "tweet length": 44,
  "links": [
      "https://t.co/5E624yD9uv"
  ],
  "translation": "",
  "pos": [ ....
  ],
  "hyponyms": {
    "Hyponyms": [ ...... ]
  },
  "named entity": [ ....
  ],
  "topic": [
    "democratic",
    "pay"
  ],
  "sentiment": [
    "1",
    "-3"
  ],
  "check_dictionary": 0.9130434782608695,
  "Named count": 202,
  "names": [],
  "html": null,
  "pure_text": ""
}
\captionof{lstlisting}{Example of tweet json format at
   the end of the pipeline after exporting to files and
    database}
```

4.1.3. Data analysis

Segmentation and part of speech tagging

Breaking the tweets into sentences and trying to get an understanding out of them is a typical method of text analysis. It is used to get understanding out of how many times the user used nouns, verbs, adverbs and other part-of-speech tags, which can give us some indications that can be helpful in our research. In figure 8, we present the findings in the total 17200 tweets ,from March, June and July only, related to cancer using Stanford part of speech tagger.

Getting Hyponyms and acronyms

Extracting hyponyms and acronyms is helpful so when we start matching topics it can identify the topic easier, so as the topic extracting will be explained in the next section and will give a clear understand why we used this approach. The used libraries to get hyponyms and acronyms are based Wordnet.

Named entity recognition

The point of this phase is to get how named entities out of the tweets detected so as ‘PERSON’ means that the tweet contained a mention of someone’s name, ‘LOCATION’ means it was able to get a recognized location, recognized ‘ORGANIZATION’, ’DATE’, ‘TIME’, ‘PERCENT’ or ‘MONEY’. The used named entity recognition library is Stanford.

Entity relation detection

After detecting named entities, we were able to build entity relations in the sentence between text. Using Stanford coreNLP we were able to construct parser tree of the sentence as in figure 10. This approach will help us build relations in the tweet and understand the text better to make a meaning out of the sentences.

Genia Tagger for medical text

Genia Tagger is used part-of-speech tagging, shallow parsing, and named entity recognition for bio-medical text. The use of Genia Tagger was to test whether it will be useful for tweets as twitter is only a source of unstructured data that does not contain medical text. So, after testing this tool on twitter as one of the biggest social media platform, we think this tool will not be helpful for such platforms and it requires medical text to give more knowledge.

Topic detection

Topics detection is extracted using Latent-Dirichlet Allocation (LDA), which is a topic model and was first presented as a graphical model for topic discovery. Using this approach to extract topics out of tweets gave us the results which shows how many times the topics were extracted from the text, these topics are then matched with keywords related to family, money, friends or treatment keywords to classify the tweets under major branches to give more indication of how people use twitter and other social media platforms to talk about cancer issues.

Emotion extraction

The final stage and the target from the project is to get emotional meanings out of these tweets or text, sentiment analysis tools are different and we are using IBM Watson natural language understanding to extract emotional meanings from the text. The results were under examination and it can give us more understanding and in depth meanings of the tweets. But, IBM Watson was not convenient in our case, it has limited usage of words and characters. We had to look for alternatives. The best one was SentiStrength which estimates the strength of positive and negative sentiment in short texts, even for informal language. It has human-level accuracy for short social web texts in English, except political texts. SentiStrength reports two sentiment strengths: (-1 to -5) for negative sentiment and (1 to 5) for positive sentiment.

4.1.4. Psychosocial categorisation

The last step was to build a categorisation model based on the keywords and the topics extracted from the LDA phase and the Hyponyms matching, populating them together and comparing them with a predefined made list for four basic categories: family, friends, treatment and money related keywords. We were able to catch significant amount of tweets that matched with these categories which explains how much of the topics extracted from these tweets matched with each category, noting that some of these topics might exist in many categories from the following, but this gives us indication how much tweets are related to each other and then we can build more improved classifiers. Figure 3 explains how that model works. We choose five different categories; family, friends, money, treatment and lifestyle. Family related issues are where most of the communication and family related issues should be contained in it. Friends are usually the comfort where patients tend to. Money is related to the financial problems that patients might face while treatment is about treatment challenges. Finally, lifestyle might be affected by the disease so it will good to examine such changes.

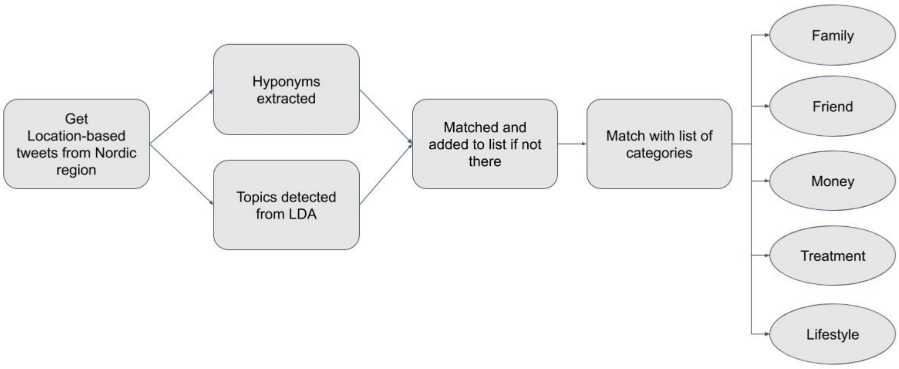

Get Location-based tweets from Nordic region
Hyponyms extracted
Topics detected from LDA
Matched and added to list if not there
Match with list of categories
Family
Friend
Money
Treatment
Lifestyle

Figure 3. Cancer psychosocial categorisation model

4.1.5. Cancer types categorisation

The second analysis point is trying to identify the types of cancer extract from the tweet, as mentioned in Cancer.gov, a central website for the National Cancer Institute (NCI), the U.S. government's principal agency for cancer research, that for every cancer type there should be some keywords that can help identify that this tweet relates to a specific type of cancer. 'Tract cancer' and 'gastrointestinal' can relate to stomach cancer while 'LCIS' is usually used in the breast cancer domain. We classified the types of cancer into 9 types suggested and with similar keywords as explained previously, these 9 types where:

- Stomach
- Breast
- Skin
- Bone

- Pediatric (Childhood cancer which usually involves kids and keywords related to such age is included in this type as well).
- Brain
- Head and neck
- Blood, also referred to it as Leukemia.
- Lung

The analysis in this part was done on larger data-set including Streaming tweets from March, June and July in the Nordic region.

Keras is a deep learning library to detect age and gender from pictures of people. The model was trained on different pictures from IMDB and other sources as a training material. The model accuracy for gender is 51% while for age is 10%, which indicates the need to improve the model in age detection. However, the model is still under testing and development, by adding this model, we will be able to get verified values about gender and age detection from twitter users who write about cancer in the Nordic region or outside.

We were not able to use Keras, however, as it is not easy to use it on Twitter profile pictures. Such pictures might not have clear images or even no images at all. This was a challenge affecting the performance and accuracy of such tool, so instead we used keyword matching approach as the rest of the project. Keyword matching approach is based on the assumption that when a Twitter user mentions "my wife", for example, it is a high probability that this person might be male. This approach is followed on the rest of the project, so it will be consistent to use instead for such a case.

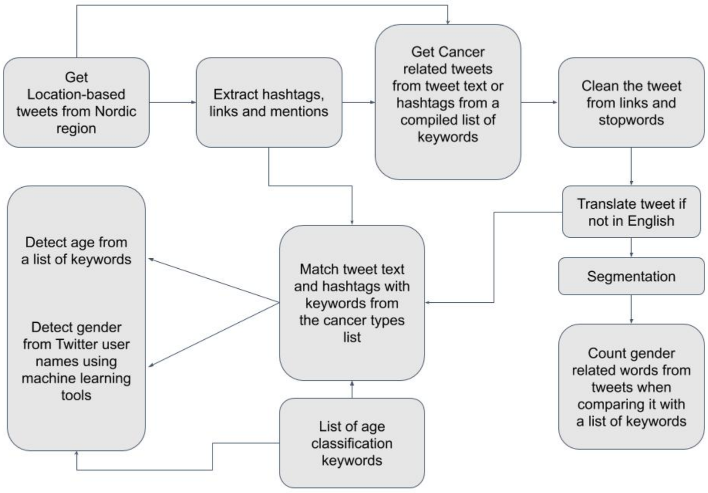

Get Location-based tweets from Nordic region
Extract hashtags, links and mentions
Get Cancer related tweets from tweet text or hashtags from a compiled list of keywords
Clean the tweet from links and stopwords
Translate tweet if not in English
Segmentation
Count gender related words from tweets when comparing it with a list of keywords
Match tweet text and hashtags with keywords from the cancer types list
Detect age from a list of keywords
Detect gender from Twitter user names using machine learning tools
List of age classification keywords

Figure 4. Cancer type categorisation model

Algorithm 2 cancer type, age and gender detection

```
Input   list of cancer related keywords: Cancer[], list of keyword for every
cancer type: TypesLists[], lists of keywords for age detection: AgeLists[]
and gender detection: Male[] and Female[], outputDir, Pretrained model on
gender detection from usernames
Output DataFrames and percentages.
  procedure CancerType
    while longitude and latitude in Nordic Region do Save Tweets in JsonFormattedFiles in
  outputDir;
    end while
    for every file in outputDir do
      for every Tweet in file do
        Extract and remove Hashtags, links and mentions;
        Username ←Extract username from file;
        if any of k[] in Tweet or Hashtags then
          Get language from Tweet;
          TranslatedTweet ←translate Tweet to English;
          Sentences ←Divide sentences in TranslatedTweet
          for every word in sentence do
            if word in Male[] then
              INCREMENT MaleWords;
            else if word in Female[] then
              INCREMENT FemaleWords;
            end if
            function ReturnGender
              if MaleWords > FemaleWords then
                Percent ←(MaleWords/LengthOfTweet)*100;
                return  Percent, Male
              else if MaleWords < FemaleWords then
                Percent ←(FemaleWords/LengthOfTweet)*100;
                return Percent , Female
              else
                return Both
              end if
            end function
          end for
          Sentences ← Divide sentences in TranslatedTweet;
          Tokens ←Tokenize words in Sentences;
          for every Type in TypeLists[] do
            if Tweet contains a word in TypesLists[] then
              INCREMENT Type;
              for every Type in TypeLists[] do
                if every Age in AgeLists[] then
                  INCREMENT Age;
                end if
              end for
              Username ←Get part of speech tags;
              Gender ←Get named entity recognition;
              return Age, gender;
            end if
          end for
        end if
      end for
    end for
    Dataframes ←Age, gender, Percent, Type
  end procedure
```

4.1.6. Cancer stages categorisation

Defining the stage of the cancer disease was made previously using machine learning tools as SVM and medical data that was divided easily into training and evaluation data set. However, the model required some annotated and defined text previously which is not available in our approach, although it can be used to improve the detection but twitter text is so small to be able to do that [15]. So, we followed another approach as illustrated in the next figure, keywords matching and the streaming tweets are used as our data-set, then by extracting hash-tags, links and mentions, we were able to catch tweets related to cancer and then clean these tweets, translate them and prepare them for further analysis to be classified based on keywords found in the tweet text. To classify based on the stages of cancer, we used two approaches to catch the stages the first one had the following stages (Stage 0 - Stage 1 - Stage 2 - Stage 3), where keywords related to these stages are included per each. For example. Stage 0 can include "benign" and "cancer" keywords as it is the first stage of tumour identification. Other approach is catching the TNM factors and numbers, which can give more indication on the cancer stage from these numbers. The total data set combines of total 17200 tweets, which contains keywords included in our sets. TNM staging system is a globally recognised system for classifying the extent of spread of cancer. TNM is a notation system that describes the stage of a cancer, which originates from a solid tumor, using alphanumeric codes:

- **T**: describes the size of the original (primary) tumour and whether it has invaded nearby tissue
- **N**: describes nearby (regional) lymph nodes that are involved
- **M**: describes distant metastasis (spread of cancer from one part of the body to another).

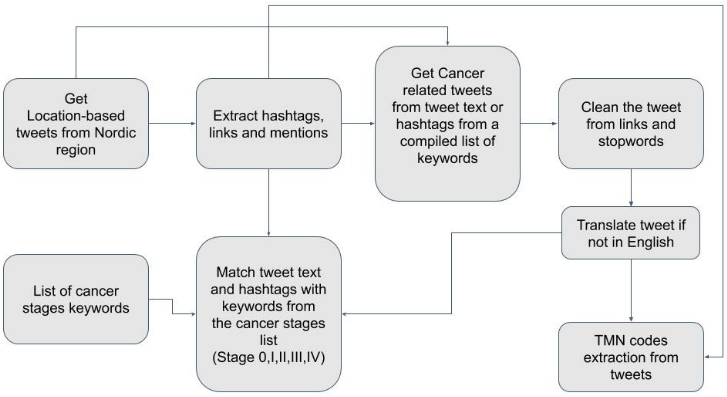

Get Location-based tweets from Nordic region
Extract hashtags, links and mentions
Get Cancer related tweets from tweet text or hashtags from a compiled list of keywords
Clean the tweet from links and stopwords
Translate tweet if not in English
List of cancer stages keywords
Match tweet text and hashtags with keywords from the cancer stages list (Stage 0,I,II,III,IV)
TMN codes extraction from tweets

Figure 5. Cancer stage categorisation model

4.2. Specific Twitter user analysis

In this approach, we wanted to focus on a specific twitter user, as a sample in this section, we decided to use a specific user, who is tweeting a lot about cancer. We were able to catch this user using Twitter search API looking for the word 'cancer'. The target of this approach is to measure the psychological effects upon that user over different periods of time were disease progress. This will give us a clear understanding about how the disease progresses in even the psychological and social aspect. Building graphs per time periods is very important for that approach and analysing each part independently to get specific knowledge. The following results were fetched from a specific user that will be easy to analyse as she documented most of her troubles with cancer online and it will be helpful to build the following graphs on timescales. The processing approach is quite close to that in streaming tweets approach. Cancer type and stage were already determined from the user itself, then we didn't have to check for these from the user tweets.

4.2.1. Data collection

After applying the twitter search API, we were able to catch many users tweeting about cancer, to test our approach we used the twitter user 'Julie McCrossin' who is a cancer survivor sharing advices and personal experience with her cancer problems with a total of 23k tweets but we took a sample of 3200 tweets to conduct our experiments. She used to mention mostly cancer related accounts, surgeons and medical institutions. In one of her tweets, Julie was mentioning the help of radiation in saving her life when she said "Radiation saved my life and preserved my voice when I had throat cancer. Some patients aren't offered the option". Highlighting the need for more affordable ways to provide such treatments to cancer patients and how some might suffer to be able to cover such expenses.

4.2.2. Formatting and cleaning data

The formatting and cleaning procedure is different in this part as we are not filtering tweets in the specific user analysis as we are interested in the development of sentiment and topics which indicates the development of such factors along time. Also, in this approach, the selected user is already an English speaker twitter user.

4.2.3. Data analysis

Same as in the first approach, dividing the tweet into sentences and getting part of speech tagging from them. This will give us an indication about the user itself and how she uses the language and this will give indication about the user's personality as well. Mentions of persons, organisations and locations that the user knows and whether she uses these mentions too much or not. Emotion extraction in this approach the emotion extraction will change along the duration of tweets.

Analysis includes Segmentation and part of speech tagging where we divide the tweet into sentences and getting part of speech tagging from them. This will give us an indication about the user itself and how she uses the language and this will give indication about the user's personality as well. Named entity recognition which are mentions of persons, organisations and locations that the user knows and whether she uses these mentions too much or not. Getting Hyponyms and acronyms which are used in the categorisation model alongside topic detection models. Finally, emotion extraction will be along specific periods where we focus on the variations and changes and whether they are triggered by specific events or states.

4.2.4. Psychosocial categorisation

The last step was to build a categorisation model based on the keywords and the topics extracted from the LDA phase and the Hyponyms matching, populating them together and comparing them with a predefined made list for four basic categories: family, friends, treatment and money related keywords. We were able to catch significant amount of tweets that matched with these categories, the showed figure 3, explains how much of the topics extracted from these tweets matched with each category, noting that some of these topics might exist in many categories from the following, but this gives us indication how much tweets are related to each other and then we can build more improved classifiers. The figure explains how that model works.

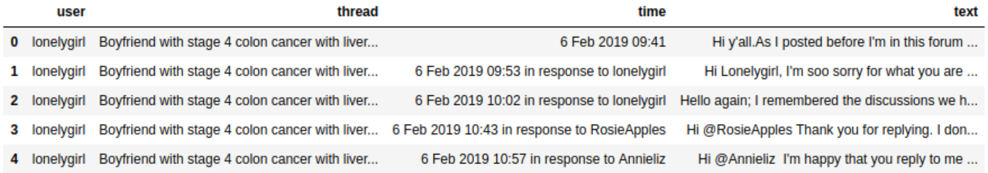
4.3. Online forums analysis

The third part of our research is to focus on a specific online platform for analysis, where we chose `cancerresearchuk.org` to do our analysis, this platform is considered one of the main platforms for cancer research, exchanging information, advises, emotional and all types of support.

4.3.1. Data collection

In the beginning of this part of analysis, we had multiple candidate websites to use for analysis. cancerUK was one of these candidates and other online forums for cancer as `onlinecommunity.cancercouncil.com.au` and others. To fetch the data needed from such websites, we had to construct our own scrapers after search tools to get the best candidate for our application, tools like scrapy, beautifulSoup and scraper API. We used beautifulSoup tool to be able to collect data from the website. But we also had to go through muliple pages in the platform as the website has many pages to scroll through, so using an open-source web-based automation tool as selenium for testing and automation.

Using both selenium and beautifulSoup enabled us to go through pages and fetch needed data which is the user name, if it is a main thread or a reply, the post text, replies, date and other important information to help our further analysis. Later, it was important to keep the data somewhere, so we will not need to run scraping scripts every time we need to do more data analysis later. Saving time in this step was essential as we learnt from previous implementations on Twitter, to optimise speed and performance more. Data was saved first on MongoDB, a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schema. But then later, this implementation was not obsolete enough to manage huge text and multiple values with it, so we changed to dataframes implementation using CSV files which provides easier, faster and consistent access, edit and organise data. The output should look like in figure 6 with a sample user, thread title, time when it was posted and text of the post [2].



	user	thread	time	text
0	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 09:41	Hi y'all.As I posted before I'm in this forum ...
1	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 09:53 in response to lonelygirl	Hi Lonelygirl, I'm soo sorry for what you are ...
2	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 10:02 in response to lonelygirl	Hello again; I remembered the discussions we h...
3	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 10:43 in response to RosieApples	Hi @RosieApples Thank you for replying. I don...
4	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 10:57 in response to Annieliz	Hi @Annieliz I'm happy that you reply to me ...

Figure 6. Sample of data after scraping

Table 4 shows the users, we detected in our analysis and the distribution of posts and cancer type. Also, whether the user himself/herself is the cancer patient or the user is doing the communication on the social media for other patients who are related. This will be interesting in our analysis, in order to identify how relative react and behave on a psychosocial level from the online forum.

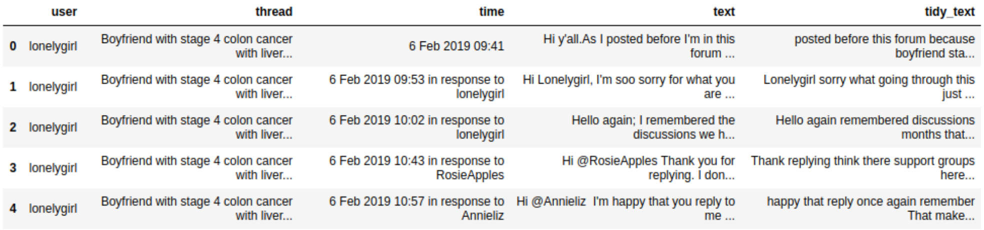
[2] Dataset can be accessed online: `https://www.kaggle.com/moamenibrahim/text-mining-and-analysis-on-cancer-uk`

Table 4. social media distribution in Nordic region

User	**Cancer details**	**Patient**	**Number of posts**
User 1	Liver and Lung metastasis	Relative to user (Boyfriend)	24
User 2	Brain cancer	Same User	70
User 3	Breast cancer	Same User	674
User 4	Throat cancer	Same User	197
User 5	Triple negative breast cancer grade 3	Same User	13
User 6	Breast cancer	Same User	146
User 7	Neuroendocrine tumors (NET) cancer	Same User	41
User 8	Breast cancer	Same User	16
User 9	Ones and lymph node cancer	Same User	15

4.3.2. Formatting and cleaning data

Target of cleaning in this part is to prepare the text for further steps in the analysis procedure. To perform proper analysis using the tools that we leveraged, we had to prepare the text in the form that will make it proper to apply our tools. For that we need to perform tokenization and stemming. Figure 7 shows how the text looks like after tidying and preparing.

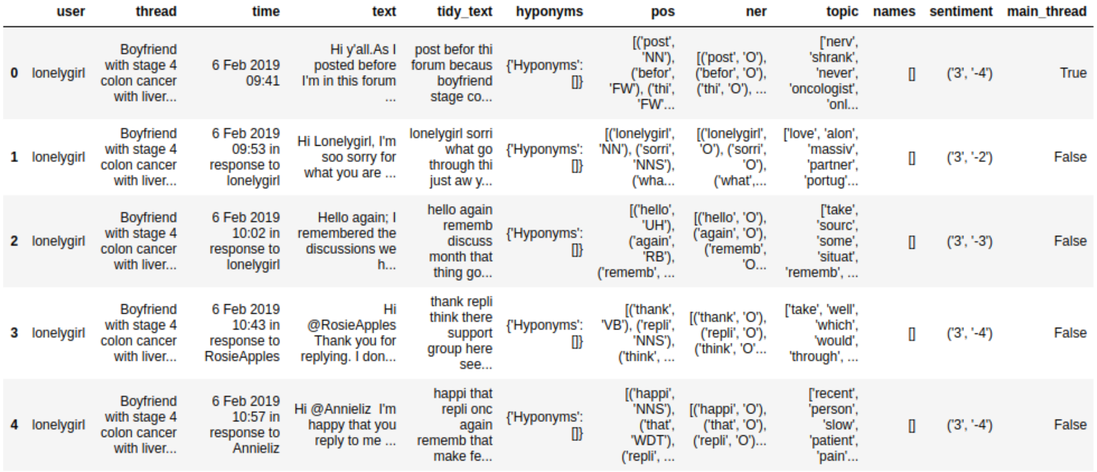


	user	thread	time	text	tidy_text
0	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 09:41	Hi y'all.As I posted before I'm in this forum ...	posted before this forum because boyfriend sta...
1	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 09:53 in response to lonelygirl	Hi Lonelygirl, I'm soo sorry for what you are ...	Lonelygirl sorry what going through this just ...
2	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 10:02 in response to lonelygirl	Hello again; I remembered the discussions we h...	Hello again remembered discussions months that...
3	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 10:43 in response to RosieApples	Hi @RosieApples Thank you for replying. I don...	Thank replying think there support groups here...
4	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 10:57 in response to Annieliz	Hi @Annieliz I'm happy that you reply to me ...	happy that reply once again remember That make...

Figure 7. Sample of data after formatting and cleaning

4.3.3. Data analysis

Data analysis part focused on the part of speech tags, named entities, topics, detected names if there was any and sentiment analysis. In this approach, part of speech tags were used to detect how sentences are formed on such platforms. Trying to get sense from the lexical analysis tools, as the tokenization, stemming and part of speech tagging. Named entities shows how common specific organisation, person name, time or date is mentioned in the text, so this will be an interesting finding in such analysis. Topics will help in the categorisation model, we defined before. On the other hand, sentiment analysis will define how likely this post is categorised on the emotional level, which of course will determine whether the post is positive or negative, and by correlating these with the topics, we will be able to identify key findings related to each topic. Figure 8 shows the output of each part of speech tags, named entities, topics, detected names if there was any and sentiment analysis applied on the text from each thread.



	user	thread	time	text	tidy_text	hyponyms	pos	ner	topic	names	sentiment	main_thread
0	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 09:41	Hi y'all.As I posted before I'm in this forum ...	post befor thi forum becaus boyfriend stage co...	{'Hyponyms': []}	[('post', 'NN'), ('befor', 'FW'), ('thi', 'FW'...	[('post', 'O'), ('befor', 'O'), ('thi', 'O'), ...	['nerv', 'shrank', 'never', 'oncologist', 'onl...	[]	('3', '-4')	True
1	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 09:53 in response to lonelygirl	Hi Lonelygirl, I'm soo sorry for what you are ...	lonelygirl sorri what go through thi just aw y...	{'Hyponyms': []}	[('lonelygirl', 'NN'), ('sorri', 'NNS'), ('wha...	[('lonelygirl', 'O'), ('sorri', 'O'), ('what',...	['love', 'alon', 'massiv', 'partner', 'portug'...	[]	('3', '-2')	False
2	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 10:02 in response to lonelygirl	Hello again; I remembered the discussions we h...	hello again rememb discuss month that thing go...	{'Hyponyms': []}	[('hello', 'UH'), ('again', 'RB'), ('rememb', ...	[('hello', 'O'), ('again', 'O'), ('rememb', 'O...	['take', 'sourc', 'some', 'situat', 'rememb', ...	[]	('3', '-3')	False
3	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 10:43 in response to RosieApples	Hi @RosieApples Thank you for replying. I don...	thank repli think there support group here see...	{'Hyponyms': []}	[('thank', 'VB'), ('repli', 'NNS'), ('think', ...	[('thank', 'O'), ('repli', 'O'), ('think', 'O'...	['take', 'well', 'which', 'would', 'through', ...	[]	('3', '-4')	False
4	lonelygirl	Boyfriend with stage 4 colon cancer with liver...	6 Feb 2019 10:57 in response to Annieliz	Hi @Annieliz I'm happy that you reply to me ...	happi that repli onc again rememb that make fe...	{'Hyponyms': []}	[('happi', 'NNS'), ('that', 'WDT'), ('repli', ...	[('happi', 'O'), ('that', 'O'), ('repli', 'O')...	['recent', 'person', 'slow', 'patient', 'pain'...	[]	('3', '-4')	False

Figure 8. Sample of data after analysis

4.3.4. Cancer types and stages categorisation

Using the same implementation tools in the previous sections, we were able to catch cancer types and stages from forums, using keywords matching to identify most common cancer type or stage in the post thread. Then based on these findings we can plot the common cancer type and stage detected.

4.4. Website application

The website shows the results obtained from the streaming and the user analysis methods, the website also provide an application currently to search the Firebase database for certain queries, and options for both Twitter and Suomi24 databases. For instance, users to the website can use it to search for English tweets related to cancer or search the database for tweets containing a specific keyword or element. The used framework is express on Node.JS which is a light-weight web application framework to help organise the web application on the server side. Also, it can support API calls to web pages and also to the search option. Express.JS basically helps managing everything, from routes, to handling requests and views. The website is built as a REST API on http GET, POST methods only to fetch pages and its values, also to create a query to be passed to Firebase. The following figure shows the architecture of the website. [3]

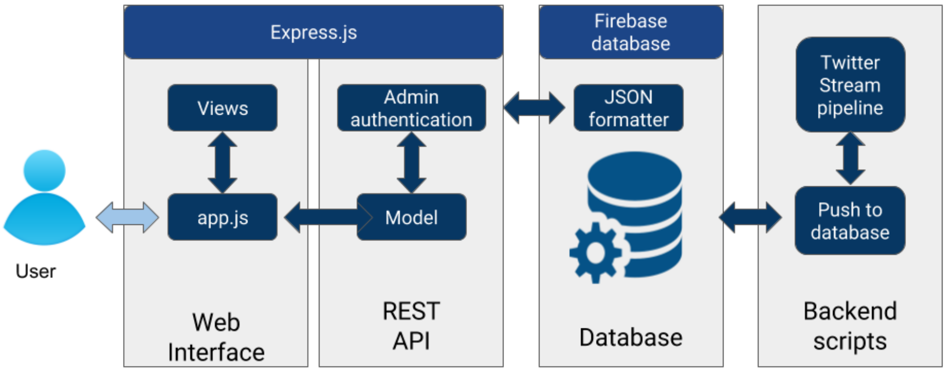


Figure 9. Website architecture

[3] Revealing effects of psychosocial factors in cancer from ESKO database and social media website can be accessed through `http://18.185.118.17:10002/`

5. RESULTS

5.1. Twitter analysis on Nordic

First approach required a proper filtering in the beginning, with the existence of many tweets in different languages as Finnish, Swedish, Norwegian and English. We used keywords matching in many different languages in order to fetch cancer related tweets only. Original count of tweets were 1.6 million tweets in the Nordic area. While the cancer related tweets were only 30 thousand, making 1.1% of the total number of tweets. These tweets are used in our analysis, whereas the following results are obtained from these 30 thousand tweets.

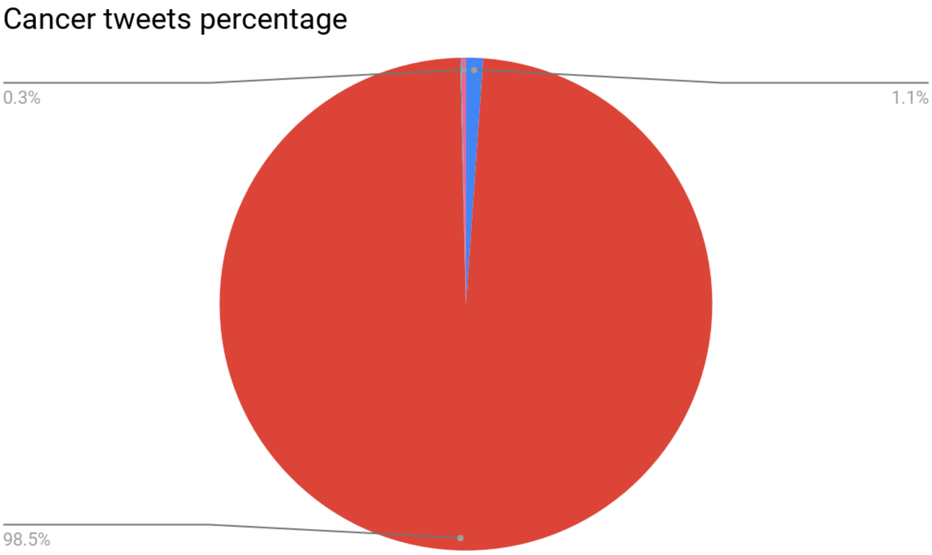


Figure 10. Percentage of tweets related to cancer

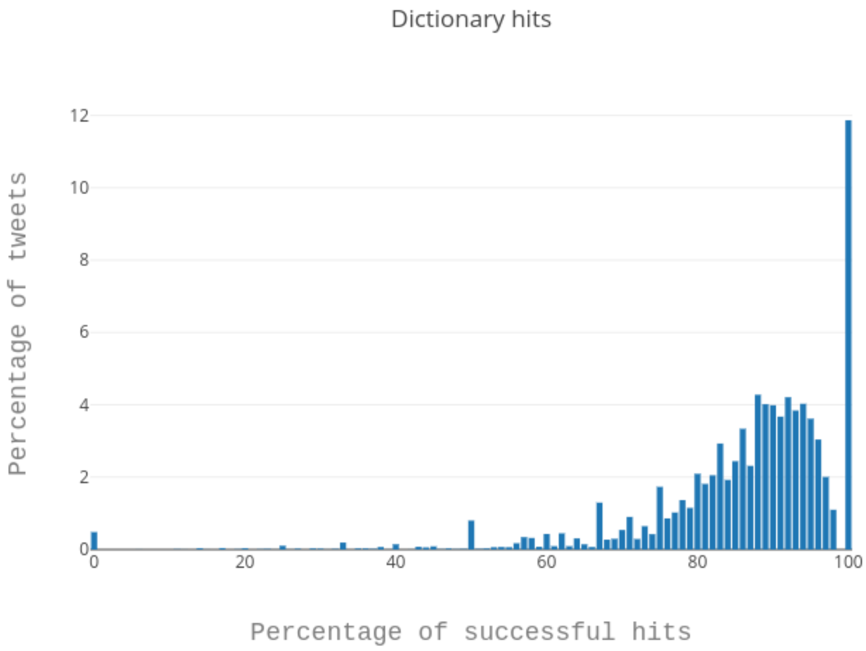


Figure 11. Y-axis: number of tweets, X-axis: Percentage of successful hits in the dictionary

Next stage was to translate tweets in other languages to English, as most of the tools used in our analysis support English better than other languages as Finnish, which is rare to find resources for. Figure 11 shows the accuracy of translating the tweets from other languages to English. As example, Finnish language has the dependency parser developed by University of Turku, which can provide parsing for Finnish text and we used it on one of the tweets in Finnish as shown in figure 12.

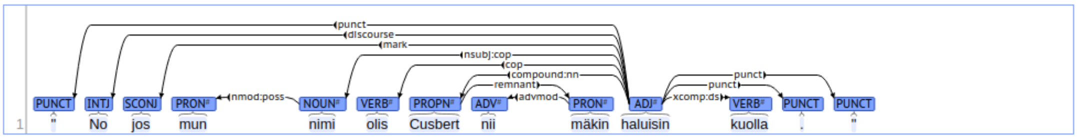


Figure 12. Sample output of the Finnish parser

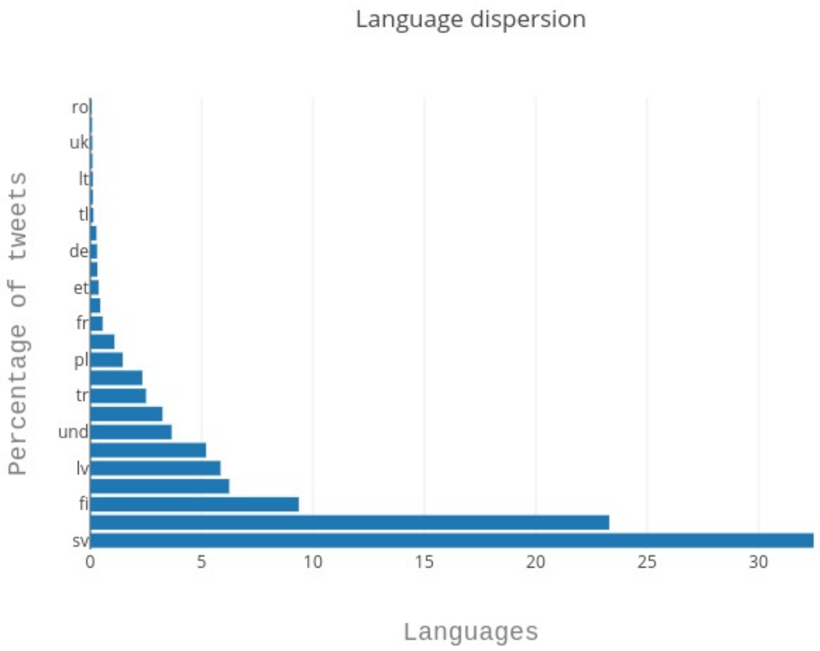


Figure 13. Language distribution in the Nordic region

In figure 13, the graph shows the popular languages found in the Nordic region, where ‘fi’ stands for Finnish, ‘en’ is English, ‘no’ is Norwegian, ‘sv’ is Swedish , ‘da’ is Danish, ‘fr’ is French, ‘ru’ is Russian and ‘und’ is undefined language for the tool used. Swedish appears to exceed all languages at that region, which can give us indication about what language many users in that region speaks and where they can be from.

Figure 14 shows the number of tweets with their length as number of words in these tweets. It shows that most of the tweets were between 0 to 50 words, indicating that tweets were mostly short and less amount of tweets were extended and had more text and probably more information that can support the findings.

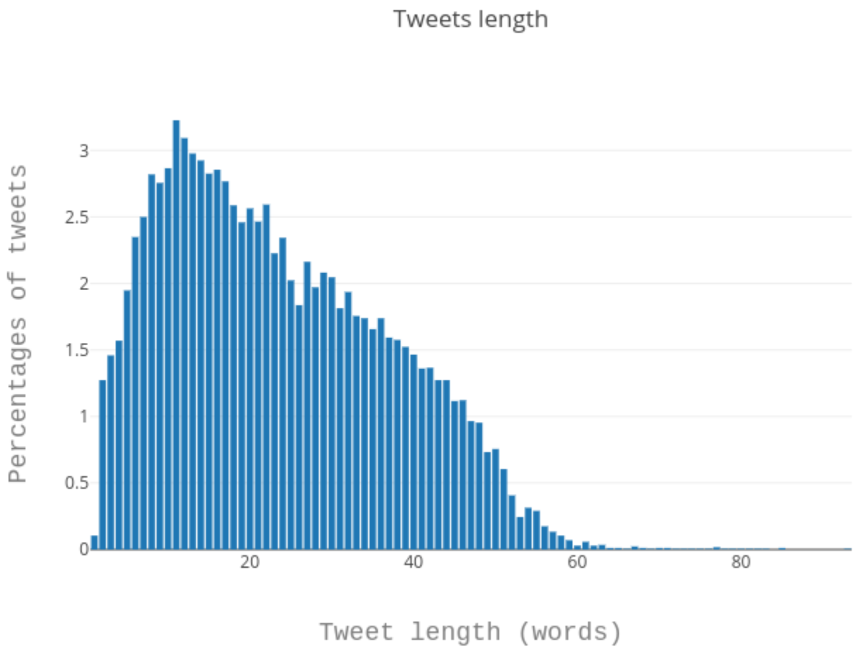


Figure 14. Words per tweet frequency

In figure 15, we present the findings in the total cancer related tweets ,from March, June and July only, related to cancer using Stanford part of speech tagger. Where 'NN' is a noun, making the majority of occurrence, indicates that users tend to use nouns when talking about cancer related topics. It will be an interesting finding to identify these nouns and doing more analysis to relate them depending on their named entities. We used a list of part-of-speech tags used in the Penn Treebank Project, as in the figure. Most of the abbreviations are related to these part-of-speech tags as 'NNS' is a proper singular noun and 'NNP' is a proper plural noun. Adverbs and personal pronouns are also highly used in the Nordic region. Foreign words seems to be less in these text, it can mostly relate to the translation effectiveness.

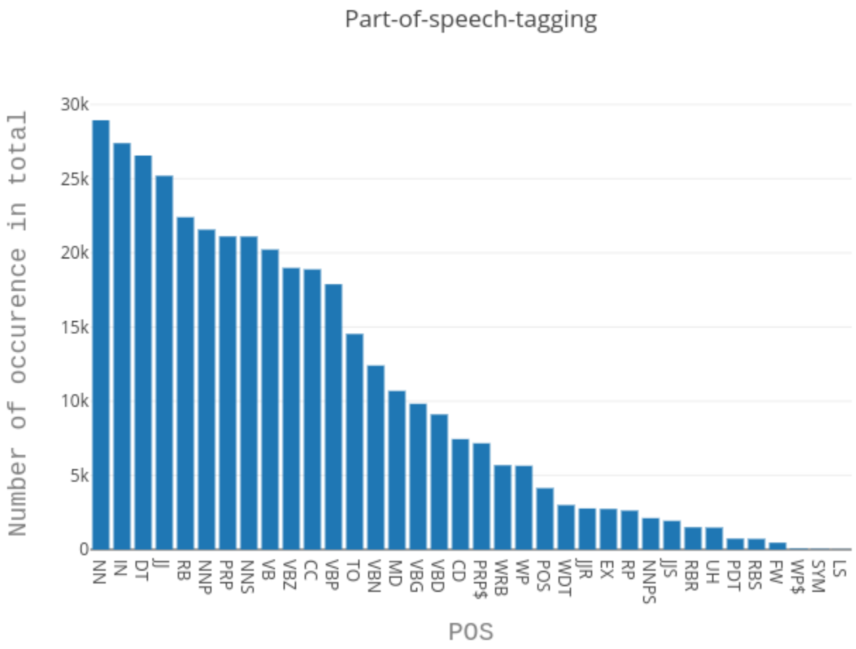


Figure 15. Part-of-Speech tagging results

Figure 16 shows how many times there was a named entity detected in the tweet, some tweets now can go up to 280 characters, while in the figure it shows named entities detected to that number and even higher, this most likely an error in the tools used as named entity count per tweet can never reach that number. But the figure shows that the more percentage of tweets are with low number of named entities meaning the relationship can be inverse proportionally. Most used named entity is location, which can be hospital, city, or clinics. It was interesting to detect in some tweets in Finland that users sometimes have to go to other hospitals in different cities than theirs in order to get proper cancer treatment. This might indicate a non-proper organisation of such service around places in Finland, which can be a challenge for patients.

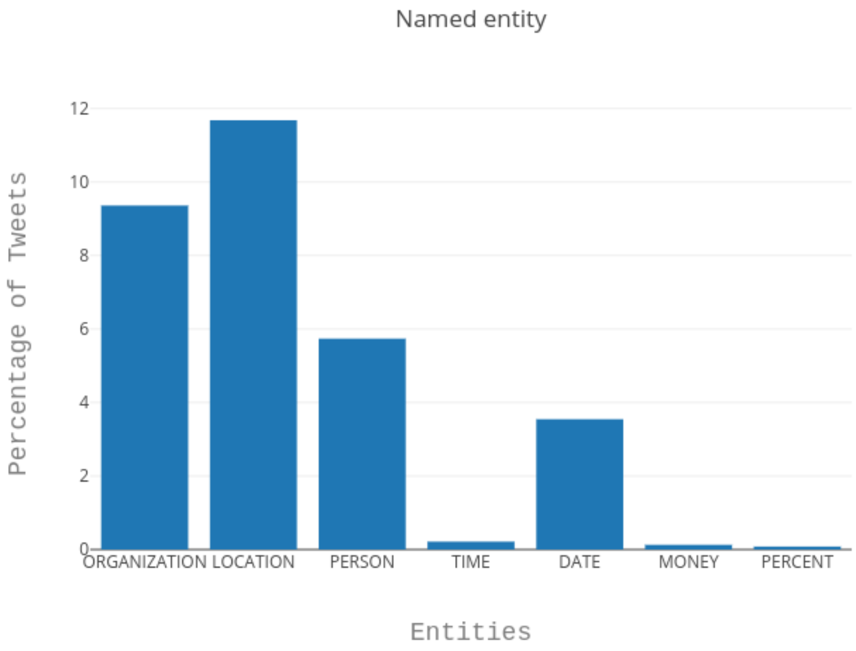


Figure 16. Percentage of tweets having these named entities

Using previously mentioned approach to extract topics out of tweets gave us the results indicated in figure 17, which shows how many times the topics were extracted from the text, these topics are then matched with keywords related to family, money, friends or treatment keywords to classify the tweets under major branches to give more indication of how people use twitter and other social media platforms to talk about cancer issues. Party, people, power and democracy were detected in many tweets, it can indicate some relation between the disease and politics. Also, a great effect for that was some political events around Finland and other Nordic countries during the time of collecting the tweets. Treatment also seem to be on the list of the highest detected topics in the Nordic area, but also a directly related topic for cancer, which relates that it is the highest in cancer concerns.

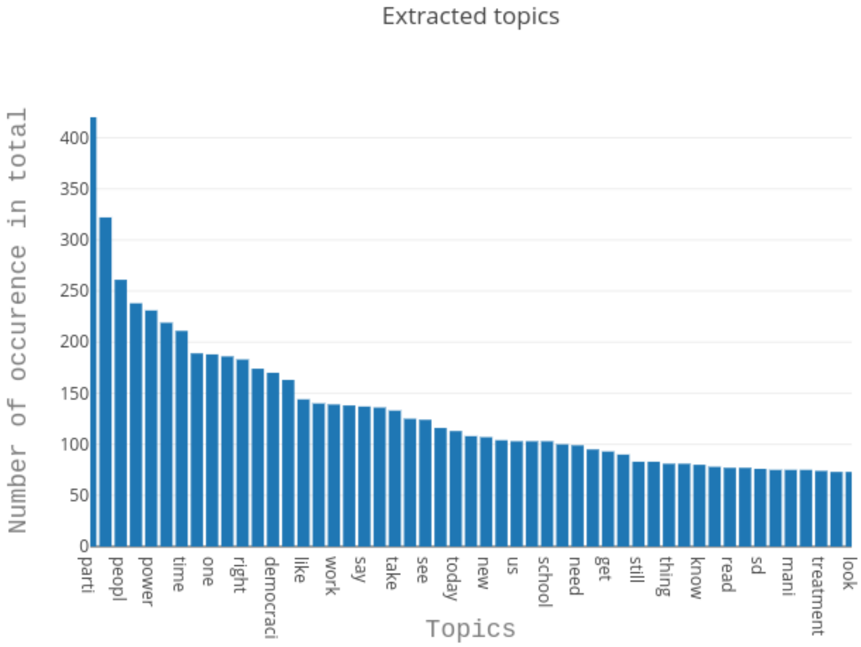


Figure 17. Topic detection results

Table 5. Sentiment analysis output on Nordic area region tweets

Sentiment score	**-5**	**-4**	**-3**	**-2**	**0**	**2**	**3**	**4**	**5**
Number of Tweets	219	1370	1518	1826	5203	2105	1792	452	64
Percentages	2%	9%	10%	13%	36%	14%	12%	3%	0%
Overall	4933				1371	4413			

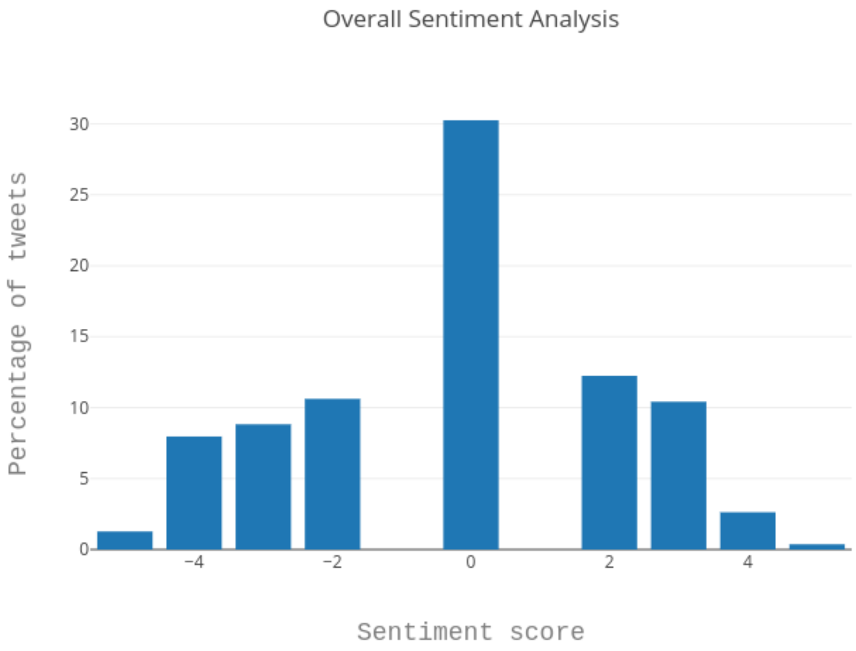


Figure 18. Sentiment analysis results

By applying the same algorithms and tools used for sentiment analysis on the categories, we were able to get the output in figure 19, it shows that treatment had more negative sentiment than most of other categories for about 7.8% of the total dataset of tweets were negative included in treatment, the same with money indicating negative emotions with such categories. Both friend and family categories showed positive emotions with 10% and 8.6% respectively, indicating some emotional support from family members and friends, this will prove that such connections should be good for cancer patients, but family category. Surprisingly, lifestyle performed well for positive emotions than that of negative and neutral emotions as shown in figure.

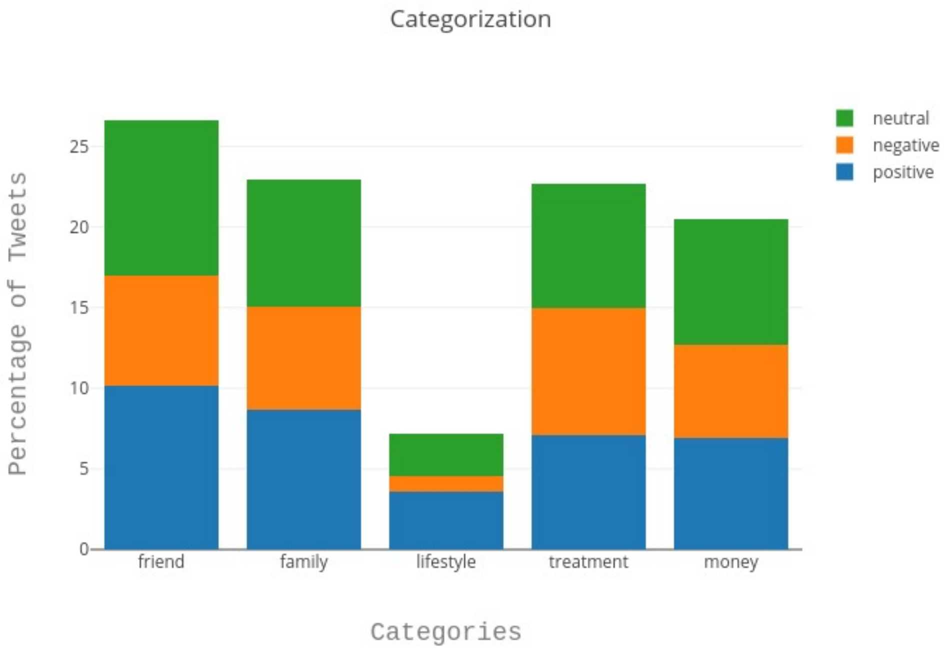


Figure 19. Psychosocial categorisation from streaming tweets

Wordcloud [44] are visual representation of the frequency of words within a given body of text. Often they are used to visualise the frequency of words within large text documents, qualitative research data, public speeches, website tags, End User License Agreements (EULAs) and unstructured data sources. Wordcloud for python was used to generate such visual output, while the most frequent words showed out to be related to these findings from the topic detection model. 'will', 'democracy' and government' indicates political topics in tweets that should be related somehow to cancer and the disease, it is somehow related to the same occasion of how people in Finland talk about political issues at the time when the tweets were collected. As in spoken Finnish, you can relate to a topic as it gives 'cancer' while the actual meaning is related to the topic itself.

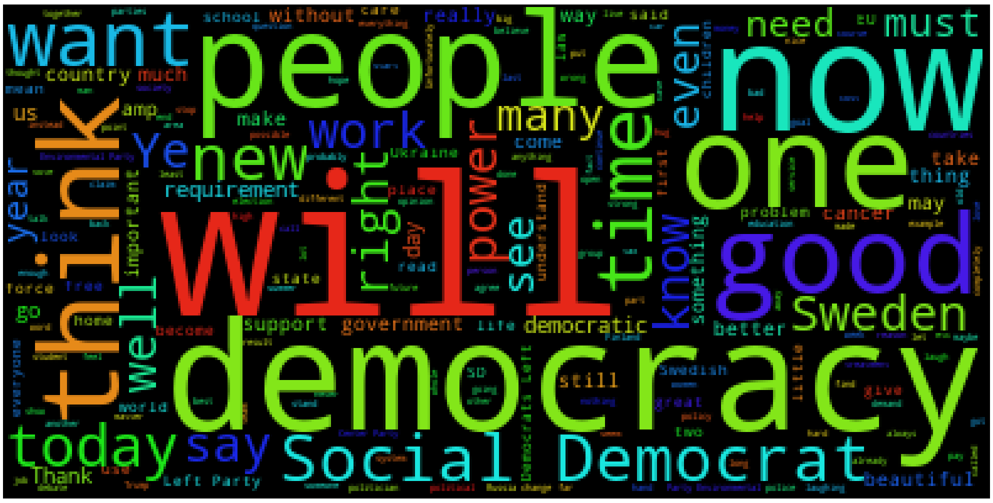


Figure 20. Wordcloud output for streaming tweets

Results out of streamed tweets from Nordic region in March, June and July 2018. The total number of tweets output is 17200 tweets. The gender distribution including the total number of tweets per cancer type is depicted in the following table with percentages.

According to the table only 2176 out of 17200 tweets (12.65% of the total data set) contained information and keywords that can indicate the cancer type.

Table 6. Cancer types gender categorisation

cancer type	**Total tweets number**	**Male number(%)**	**Female number(%)**
Stomach	233	163 (69.9%)	70 (30%)
Breast	560	355 (63.3%)	205 (36.6%)
Skin	536	312 (58.2%)	224 (41.8%)
Bone	23	11 (47.8%)	12 (52.2%)
Pediatric	364	235 (64.6%)	129 (35.4%)
Brain	57	48 (84.2%)	9 (15.8%)
Head and neck	143	106 (74.1%)	37 (25.9%)
Blood	241	174 (72.2%)	67 (27.8%)
Lung	19	7 (36.8%)	12 (63.2%)
Total	2176	1411 (64.84%)	765 (35.16%)

Age distribution was also extracted from tweets per cancer type, knowing already from previous researches that 26% of internet users aged between 18-29 years old, compared with 14% of those aged 30-49 years old.

As shown in the table, it is hard to realise the age using bag of words approach, meaning the words that these specific age ranges usually use. This made a huge challenge to detect the age from small unstructured text as tweets. The undefined age is almost 39%, making most of the tweets number. Followed by age range from 23 to 29, then 30 to 65 years old, then 13 to 18 years old making 14% surpassing the least age range which is from 19 to 22 years old. Numbers seem to verify a previous study made on Twitter users in United States but it did not focus on cancer, the study mentioned 26% for 18-29 years old age, which is close to our numbers with increase to 31.5% in cancer related tweets, while for age 30 and higher, percentage was close as it is 15.4% for cancer related and 14% normally.

Table 7. Cancer types age categorisation

Cancer Type	**Total**	**13-18 years old**	**19-22 years old**	**23-29 years old**	**30-65 years old**	**Undefined age**
Stomach	233	31	10	43	28	121
Breast	560	83	16	170	92	199
Skin	536	37	16	97	41	345
Bone	23	2	-	5	1	15
Pediatric	364	90	25	128	110	11
Brain	57	6	1	17	8	25
Head and neck	143	34	3	41	21	44
Blood	241	22	9	91	32	87
Lung	19	-	1	13	2	3
Total	2176	305 (14%)	81 (3.7%)	605 (27.8%)	335 (15.4%)	850 (39%)

Table 8. Cancer stage from Nordic area tweets

Cancer stage	**Number of tweets**
Stage 0	124
Stage 1	66
Stage 2	63
Stage 3	63
Stage 4	3

Table 9. Cancer stage from Nordic area tweets on TMN score

Cancer stage code	**Number of tweets**
T1	6
T3	3
T4	8
M1	7
M2	3
M4	3
N1	9
N2	4
N3	5
N4	3

5.2. Specific Twitter user analysis

In this approach, there is not need for translation as the selected test is already an English speaker twitter user. This is to verify the usage of Google translation API used, the same approach in the streaming tweets. Figure 21 shows tweets length for the specific user, from 10 to 20 words seems to be a common tweet length.

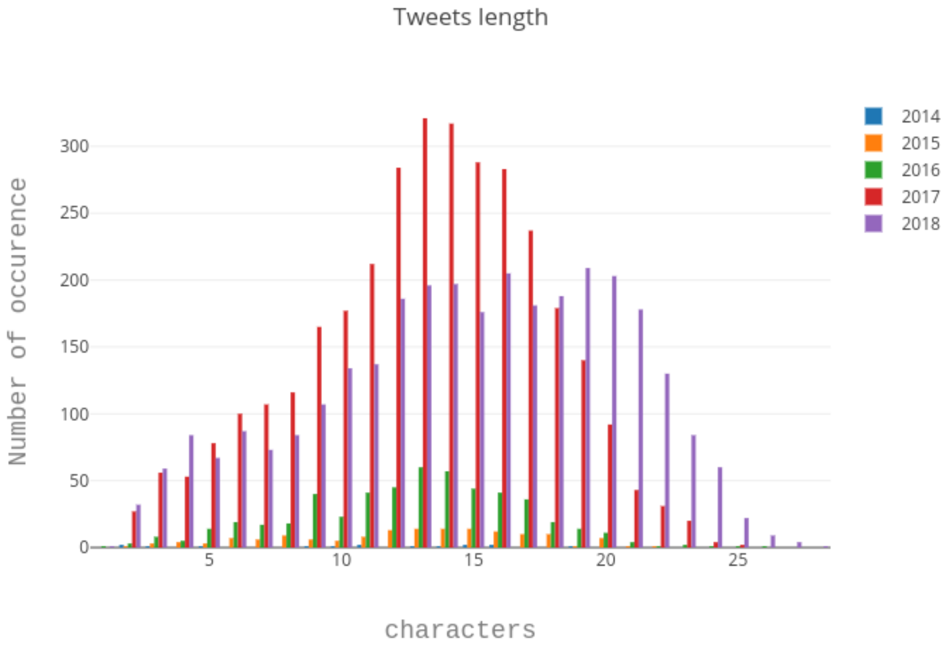


Figure 21. Tweet length for every year

Figure 22 proves the observation achieved in the previous approach as users tend to post on twitter for tweets related to cancer about people then location and lastly organisation. Least is Time, money and percent, but that can also relate to the varieties in such named entities that will make it hard to detect them.

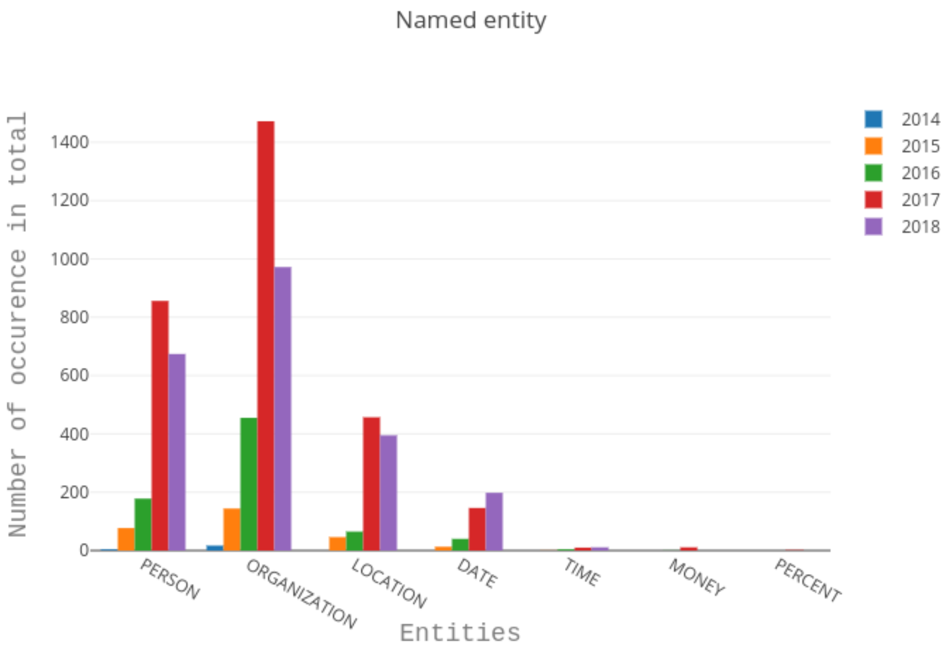


Figure 22. Named entities for every year

In this approach, we were able to detect also similarity in topics detection, as politics appeared to show up often as well in the topics. But other topics appeared as well as radiation, depression, work and other cancer related issues. This makes such approach in analysis more reliable which is what inspired more focus on similar methodology to achieve our research target.

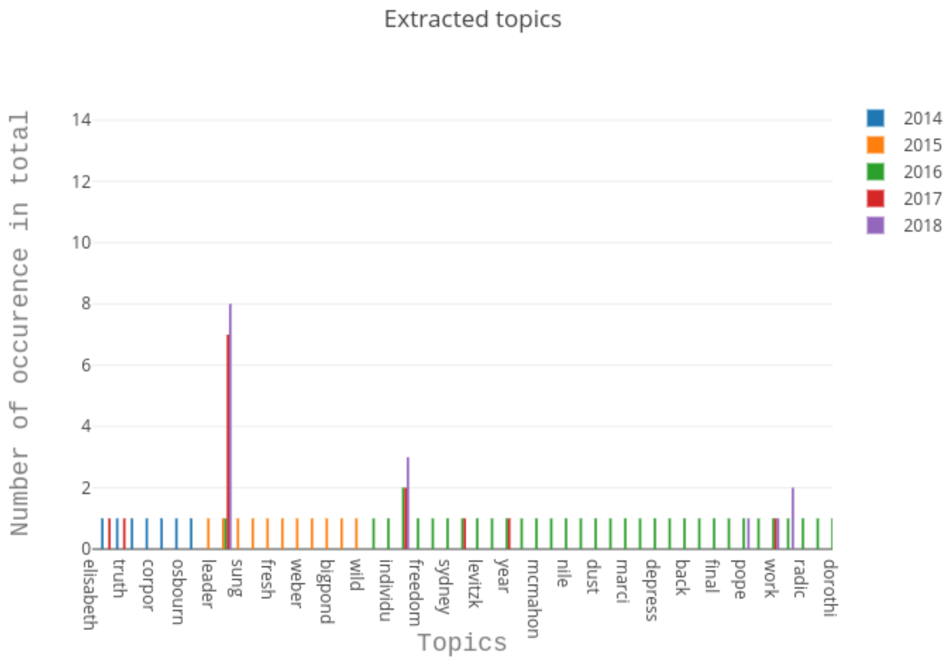


Figure 23. Topics for every year

Sentiment is mostly neutral or undetectable on user analysis making majority of tweets, while after that can be 2 or 3 positive sentiment or 4 negative sentiment. This figure 24 shows that sentiment polarity is higher in the negative scale, while in the positive scale, sentiment is positive and not much high as the counter negative one.

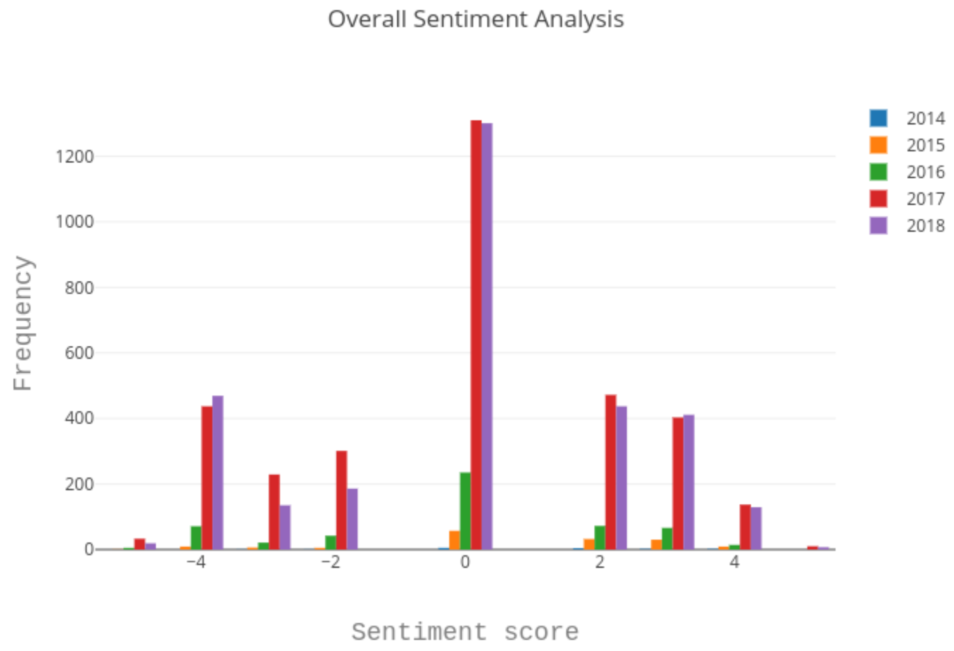


Figure 24. Sentiment per year

A difference in this analysis, than the single user analysis is that categories shows some deviation. As shown in figure 25, family and treatment were the highest, while lifestyle is the least. In comparison friendship related tweets were the highest in the Nordic area analysis. But lifestyle was also the least which shows that it is not the focus of cancer patients online.

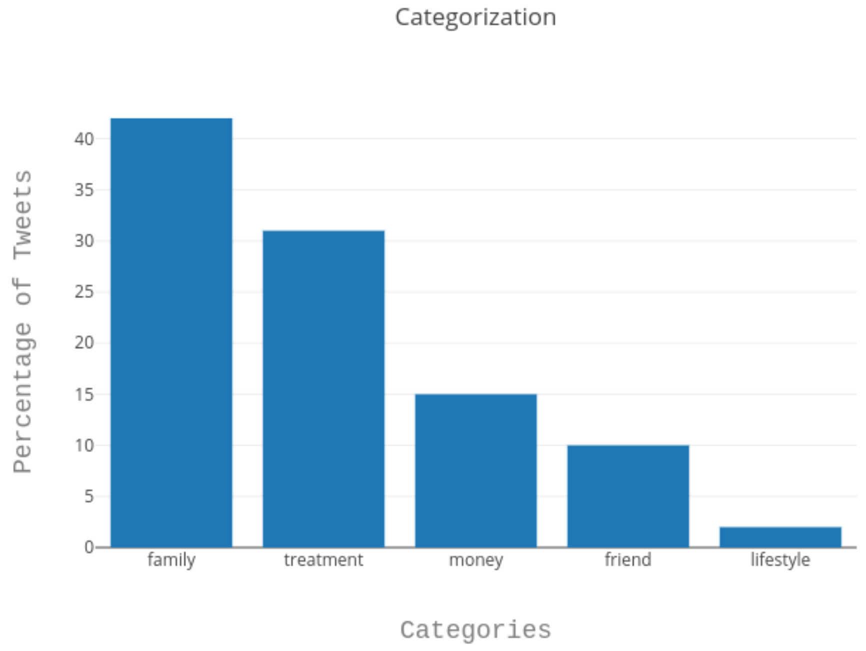


Figure 25. Total psychosocial categorisation

5.3. Online forums analysis

Last phase of analysis was the focus on online forum as cancerUK.com. In this phase, focus on specific users with lots of content was learned from other previous phases of analysis. In this approach, we focused on 9 users, where we fetched their posts, threads they replied to and their profile descriptions. Some basic information was fetched from their profiles as types of cancer they have, number of posts they interacted with or conversations started.

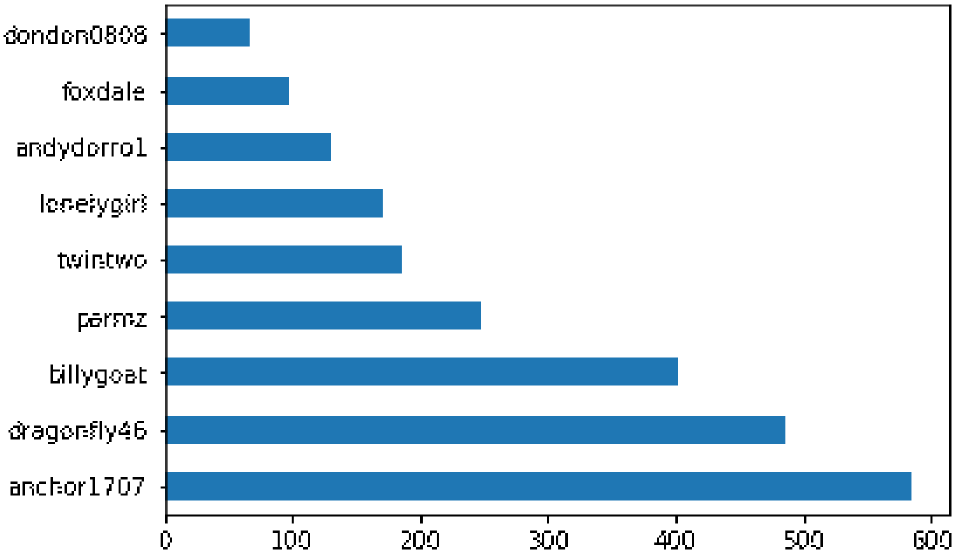


Figure 26. Number of posts per user

Figure 27 shows mostly mentioned topics from the online forum without removing stopwords. While figure 28 shows the same but after removing stopwords. It shows that topics have changed significantly, as ’cancer’, ’result’ and other treatment or disease related topics appeared more.

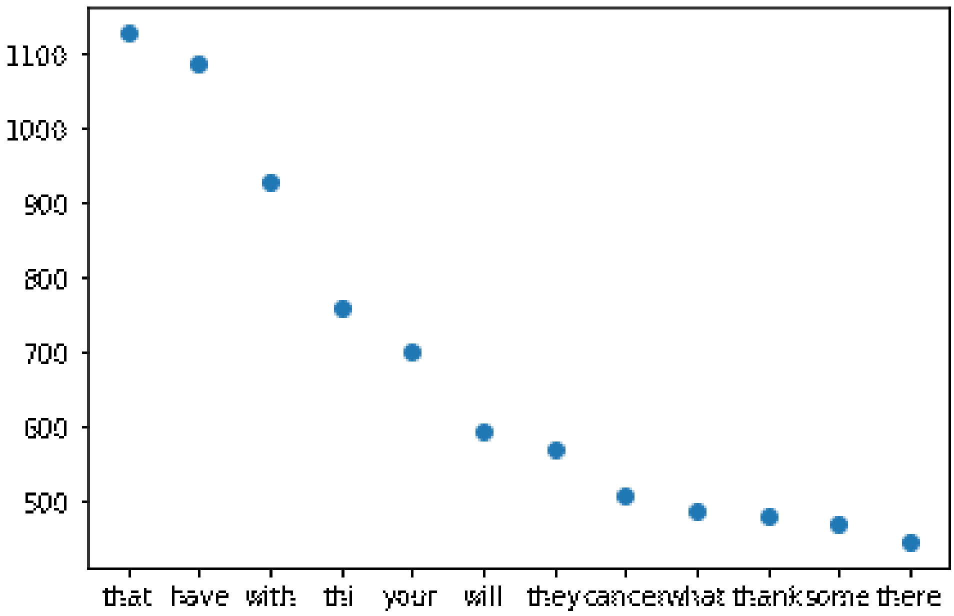


Figure 27. Most frequent topics

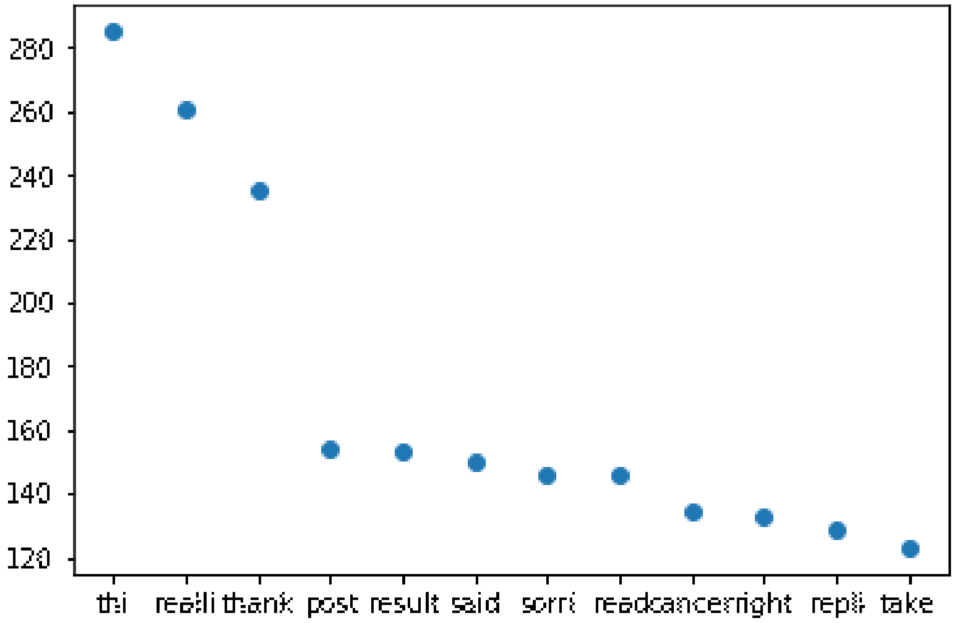


Figure 28. Most frequent topics without stop words

Figure 29 shows the word cloud dispersion output, which indicates the most repeated words among the whole text scraped from the online website. 'Treatment' seemed to be common, people talking about about treatment methods, advice and problems on these forums seems to be common, showing how cancer patients are struggling and also communicating in such matters. On the other hand, time was also common, showing patients fears and concerns. 'Help' was also a common word, cancer patients seem to seek such platform for help in information or support. Nevertheless, 'cancer', 'chemo' and other cancer related topics were repeated as words on the most common ones.

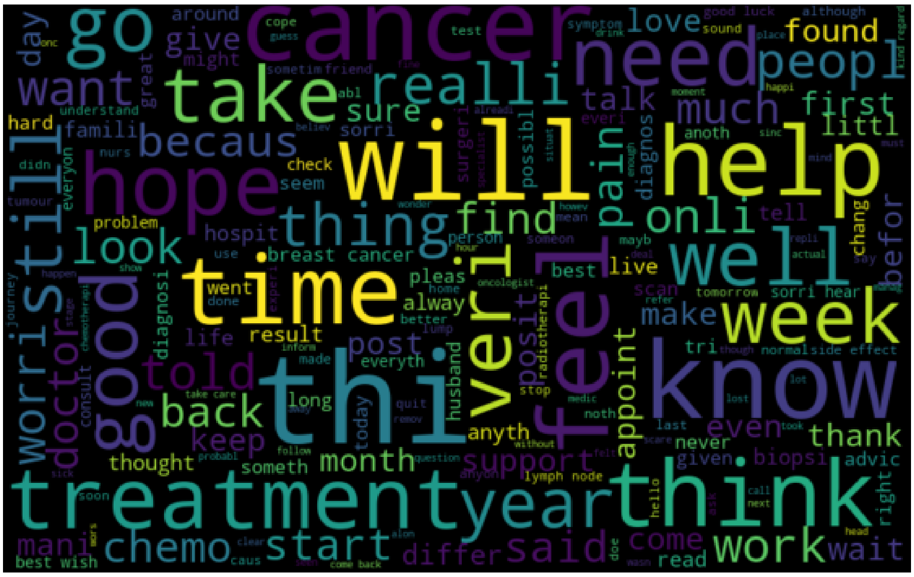


Figure 29. Wordcloud output from cancerUK online forum

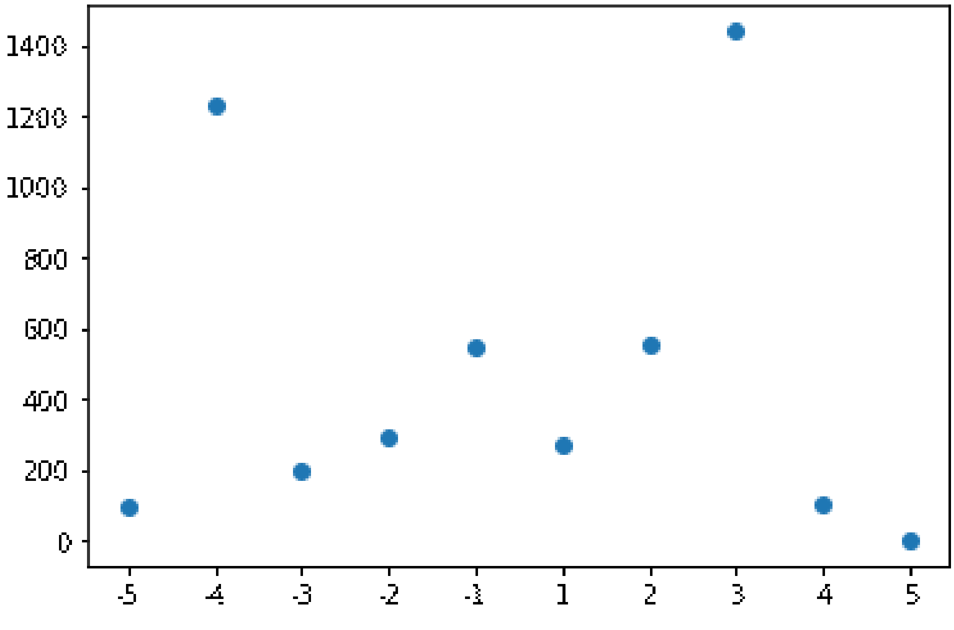


Figure 30. Sentiment score and frequency of tweets

Blood cancer seems to be topping the list in the cancer types detected on cancerUK.com, where on the contrary, breast cancer was common in the Nordic region and worldwide. Figure 31 shows the big variation between blood cancer (leukaemia) and other types of cancer as bone, brain, breast, stomach, etc.

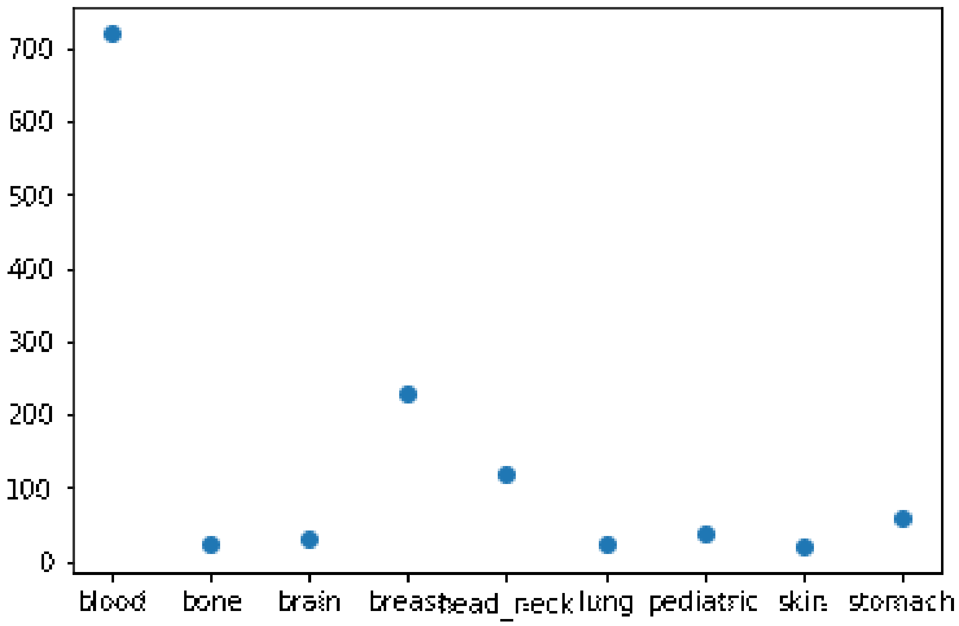


Figure 31. Cancer types detected frequency

Figure 32 shows that most of the detected cancer stages from text was stage zero, while the rest of stages were minor compared to such stage.

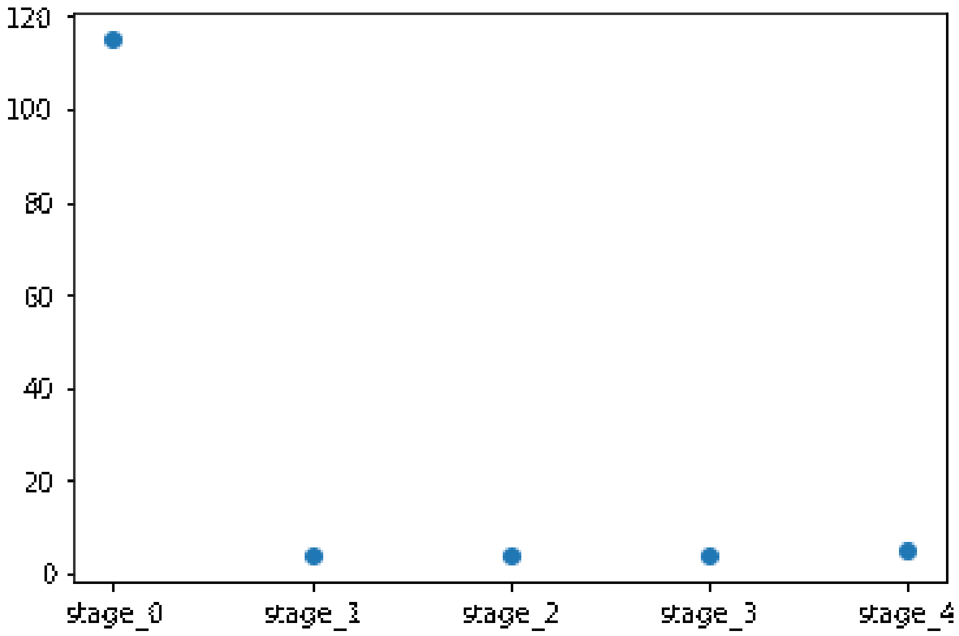


Figure 32. Cancer stages detected frequency

6. DISCUSSION

Extracting psychosocial factors from text is becoming more crucial than before, clinicians rely on such analysis in order to have proper and full treatment for cancer patients. These findings can indicate a lot about the person's experience and can show whether that person requires different care or not. It has been shown that the psychology of such patients can vary from person to another, place to another and other different classifications that mean direct or indirect impact. In this research we highlighted the importance of such topic with online forums and social media as our data source as we believe that social media and other online platforms have already became major places where users talk freely sometimes about their person experiences and also communicate thoroughly to other relatives, friends and acquaintances.

There are real barriers to this proposed social-media driven model. Patient privacy must be protected, which means implementing strict protocols for sharing confidential patient information.

In the future, it will be better to make a constructive analysis on a set of twitter users as mentioned in the first and second approach and build a visualised map showing the psychosocial and emotional tends of the user, who will be a cancer patient along a specific period of time. Comparing these results from others on different time periods or different set of users to find if there is a specific relation or pattern and get more understanding as users might talk about different topics on social media when there is a specific incident or event happening, for example.

The tools and methods used too far, analyses the text in more abstract and non constructive way. Some of these tools were conventional and could not perform well on cancer related tweets or tweets that might have emotional, social and disease related. These tools failed at some points in providing proper analysis as topic detection, sentiment, age and gender detection. All had bottlenecks and corner cases, where we were forced to use some tools with restrictions. In the future, we should be harnessing these tools and adding more to it. For example, adding social network analysis tools and building graphs that can give us more social insights about the connection and communication of cancer patients in the Nordic region (the first approach) and for the specific user (the second approach), it will be helpful to build a social network graph to know how a cancer patient or cancer survivor communicate on the social media with other people, whether they are family, friends or any other type of social connection. This will make me harness my knowledge from Social network analysis course (SNA).

Building an improved categorisation tool to classify tweets depending on the topic extracted, and the understanding of the text to get meanings out of the tweets and the social media topics. As well as knowing the context of the tweet especially for the second approach as this will build more insights how that user acts with cancer.

Age and gender detection improvement by verifying the findings from the image detection from the twitter profile picture using deep learning approach. It will verify and give more understanding about the tool and can be one step in using deep learning and machine learning approaches to build such application as we did, this will make smarter systems and improve medical and health care services, adding real time detection, help and aid to cancer patients by detecting it from social media platforms. Another machine learning outcome can be for the cancer stage detection, which will be a challenge to form a predefined and annotated text for training our models. However, this should improve our cancer stage detection tools.

Analysis from the online forum was one of the best findings we added in this research, such methodology provide reliability and good performance. Harnessing the best tools of data science, as we learned from both findings and mistakes that happened in the previous stages, we were able to use a better approach to fetch the new results. Also, we will need to expand our research to other social media platforms, not only twitter and cancerUK.com, but also other platforms as Facebook, other social media platforms and online forums.

Furthermore, doing an online program accessible by the hospital doctors and clinicians to be able to view the results, outcomes and search through the datasets we have, however, this will require input from the hospital side to give us requirements needed so they can view most of our findings and source data. We implemented an online platform, where clinicians will be able to look results from twitter and other platforms.

Finally, we identify the key findings from three different types of analysis in the following table 10 where we have pointed out the methods and their performance and reflections on each.

Points	**Streaming tweets**	**Specific user**	**Cancer UK**
Scope	Tweets in the Nordic region during a specific period of time, a challenge here is that Twitter might not be used much in this region so tweets numbers were low	Tweets for a single user only	Posts by list of specified users which might be the best approach for more analysis

Languages	Mixture of different languages which made it hard to analyse as we had to use extra translators	English as the user is English speaker, however it will be hard to compare this user to others with different languages	Perfect environment for English tools as most of the users are English speakers
Part-of speech tagging	Sentences might change structure due to translation but nouns were common	Good usage on English tweets, nouns were also common	Nouns were also common in this analysis
Named entities detection	It was good to use Stanford tools here as we were able to fetch good results after translation Location was the highest	Organisation was the highest	Different variations in named entities that can not be compared to other analysis types
Topic detection	LDA modeling on translated text might be challenging	cancer related topics were identified	Topics related to cancer were visible specifically when using Wordcloud analysis
Sentiment analysis	Sentiment might not be genuine after translation	Analysis per year should some but not major variations, however it is helpful to identify changes	Positive sentiment was a lot but the polarity of negative sentiment is high
Psychosocial categories	Friendship was the highest with more postive sentiment than negative one	Family was the highest	Not finished analysis, will be better to use it per user instead in general

Cancer type	Using gender and age classifiers we were able to find results on free text, predominant were males in specific cancer types as Blood and Stomach, in general, Breast cancer was common	Cancer type, age, gender detection was not needed as user is already identified	List of users were already identified and it was not needed
Cancer stage	First stages were always common as users communicate mostly in the start of the disease	Cancer stage detection was not needed as user is already identified	List of users were already identified and it was not needed

Table 10. Comparison between different approaches and methodologies

7. CONCLUSION

In this research, we discussed an effective and advanced way to combine different tools and software in order to create a tool that can achieve our targeted tasks, which then can give us more understandings and findings about health related topics on online forums as cancerUK.com or social media platforms as Twitter. We managed to get some findings and discussed the challenges where tools are not available much for different languages as Finnish, which developed the need to edit some tools to make it compatible with Finnish text and also discuss how the outputs were and their accuracy. We also made good use of translation tools and showed the accuracy and how efficient it is to use same language tools on text rather than translating it. We also highlighted important topics to discuss more and do further analysis related to it, as it showed significant effect on health related topics. Time analysis, also, depicts some significant changes either in sentiment analysis, disease detection and common topics, fusing these data together will also give more important findings.

After using some tests in the two approaches we explained in twitter analysis, we were able to get more insights about how cancer patients and social media users talk about cancer and this will be a good and promising point in our research. As mentioned we think that twitter is not that used in Nordic countries as other places in the world, so we might need to expand our area of research to include countries in the Nordic area or try other social media platforms alongside twitter to reinforce our findings. We applied such finding by expanding the analysis to online cancer forums such as www.canceruk.com, where we were able to get more interesting findings related to psychosocial factors and findings that can relate to it.

8. REFERENCES

[1] McGregor B.A. & Antoni M.H. (2009) Psychological intervention and health outcomes among women treated for breast cancer: a review of stress pathways and biological mediators. Brain, behavior, and immunity 23, pp. 159–166.

[2] Telepak L.C., Jensen S.E., Dodd S.M., Morgan L.S. & Pereira D.B. (2014) Psychosocial factors and mortality in women with early stage endometrial cancer. British journal of health psychology 19, pp. 737–750.

[3] Roscoe J.A., Kaufman M.E., Matteson-Rusby S.E., Palesh O.G., Ryan J.L., Kohli S., Perlis M.L. & Morrow G.R. (2007) Cancer-related fatigue and sleep disorders. The oncologist 12, pp. 35–42.

[4] Ramirez A.J., Craig T., Watson J.P., Fentiman I.S., North W. & Rubens R.D. (1989) Stress and relapse of breast cancer. Bmj 298, pp. 291–293.

[5] Lutgendorf S.K., DeGeest K., Dahmoush L., Farley D., Penedo F., Bender D., Goodheart M., Buekers T.E., Mendez L., Krueger G. et al. (2011) Social isolation is associated with elevated tumor norepinephrine in ovarian carcinoma patients. Brain, behavior, and immunity 25, pp. 250–255.

[6] Pinquart M. & Duberstein P. (2010) Depression and cancer mortality: a meta-analysis. Psychological medicine 40, pp. 1797–1810.

[7] Faller H. & Bülzebruck H. (2002) Coping and survival in lung cancer: a 10-year follow-up. American Journal of Psychiatry 159, pp. 2105–2107.

[8] Weihs K. & Politi M. (2005) Family development in the face of cancer. Handbook of Families and Health. Interdisciplinary Perspectives , pp. 3–18.

[9] Cella D., Lai J.s., Chang C.H., Peterman A. & Slavin M. (2002) Fatigue in cancer patients compared with fatigue in the general united states population. Cancer 94, pp. 528–538.

[10] Phillips K.M., Antoni M.H., Lechner S.C., Blomberg B.B., Llabre M.M., Avisar E., Glück S., DerHagopian R. & Carver C.S. (2008) Stress management intervention reduces serum cortisol and increases relaxation during treatment for nonmetastatic breast cancer. Psychosomatic medicine 70, p. 1044.

[11] Rex D.K., Johnson D.A., Anderson J.C., Schoenfeld P.S., Burke C.A. & Inadomi J.M. (2009) American college of gastroenterology guidelines for colorectal cancer screening 2008. The American journal of gastroenterology 104, p. 739.

[12] McKenna M.C., Zevon M.A., Corn B. & Rounds J. (1999) Psychosocial factors and the development of breast cancer: a meta-analysis. Health Psychology 18, p. 520.

[13] Murray H.A. (1943) Thematic apperception test. American Psychological Association .

[14] Chen D.S. & Mellman I. (2013) Oncology meets immunology: the cancer-immunity cycle. Immunity 39, pp. 1–10.

[15] Ruhlmann J., Oehr P. & Biersack H. (1999) PET in oncology: Basics and clinical application. Springer.

[16] Jacobsen P.B., Donovan K.A., Vadaparampil S.T. & Small B.J. (2007) Systematic review and meta-analysis of psychological and activity-based interventions for cancer-related fatigue. Health Psychology 26, p. 660.

[17] Fox B.H. (1978) Premorbid psychological factors as related to cancer incidence. Journal of Behavioral Medicine 1, pp. 45–133.

[18] Jensen K., Soguero-Ruiz C., Mikalsen K.O., Lindsetmo R.O., Kouskoumvekaki I., Girolami M., Skrovseth S.O. & Augestad K.M. (2017) Analysis of free text in electronic health records for identification of cancer patient trajectories. Scientific reports 7, p. 46226.

[19] Spasić I., Livsey J., Keane J.A. & Nenadić G. (2014) Text mining of cancer-related information: review of current status and future directions. International journal of medical informatics 83, pp. 605–623.

[20] Baker S., Ali I., Silins I., Pyysalo S., Guo Y., Högberg J., Stenius U. & Korhonen A. (2017) Cancer hallmarks analytics tool (chat): a text mining approach to organize and evaluate scientific literature on cancer. Bioinformatics 33, pp. 3973–3981.

[21] Duggan M.A., Anderson W.F., Altekruse S., Penberthy L. & Sherman M.E. (2016) The surveillance, epidemiology and end results (seer) program and pathology: towards strengthening the critical relationship. The American journal of surgical pathology 40, p. e94.

[22] Naslund J., Aschbrenner K., Marsch L. & Bartels S. (2016) The future of mental health care: peer-to-peer support and social media. Epidemiology and psychiatric sciences 25, pp. 113–122.

[23] Kim A.E., Hansen H.M., Murphy J., Richards A.K., Duke J. & Allen J.A. (2013) Methodological considerations in analyzing twitter data. Journal of the National Cancer Institute Monographs 2013, pp. 140–146.

[24] Sriram B., Fuhry D., Demir E., Ferhatosmanoglu H. & Demirbas M. (2010) Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 841–842.

[25] Sakaki T., Okazaki M. & Matsuo Y. (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web, ACM, pp. 851–860.

[26] Murthy D. & Eldredge M. (2016) Who tweets about cancer? an analysis of cancer-related tweets in the usa. Digital health 2, p. 2055207616657670.

[27] Chung D.S. & Kim S. (2008) Blogging activity among cancer patients and their companions: Uses, gratifications, and predictors of outcomes. Journal of the American Society for Information Science and Technology 59, pp. 297–306.

[28] Ventola C.L. (2014) Social media and health care professionals: benefits, risks, and best practices. Pharmacy and Therapeutics 39, p. 491.

[29] Ortega J.L. & Aguillo I.F. (2008) Visualization of the nordic academic web: Link analysis using social network tools. Information Processing & Management 44, pp. 1624–1633.

[30] Koskan A., Klasko L., Davis S.N., Gwede C.K., Wells K.J., Kumar A., Lopez N. & Meade C.D. (2014) Use and taxonomy of social media in cancer-related research: a systematic review. American journal of public health 104, pp. e20–e37.

[31] novoseltseva E. (2017), Natural language processing projects and startups to watch in 2018. URL: https://apiumhub.com/tech-blog-barcelona/natural-language-processing-projects/.

[32] Johnsen M., Improving natural language processing algorithm in search engines. URL: https://apiumhub.com/tech-blog-barcelona/natural-language-processing-projects/.

[33] Bird S., Klein E. & Loper E. (2009), Nltk book.

[34] De Marneffe M.C., MacCartney B., Manning C.D. et al. (2006) Generating typed dependency parses from phrase structure parses. In: Lrec, vol. 6, vol. 6, pp. 449–454.

[35] Al-Rfou R., Perozzi B. & Skiena S. (2013) Polyglot: Distributed word representations for multilingual nlp. arXiv preprint arXiv:1307.1662 .

[36] Graham S., Weingart S. & Milligan I. (2012) Getting started with topic modeling and mallet. Tech. rep., The Editorial Board of the Programming Historian.

[37] Khosrovian K., Pfahl D. & Garousi V. (2008) Gensim 2.0: a customizable process simulation model for software process evaluation. In: International Conference on Software Process, Springer, pp. 294–306.

[38] Cunningham H. (2002) Gate, a general architecture for text engineering. Computers and the Humanities 36, pp. 223–254.

[39] Kanerva J., Ginter F., Miekka N., Leino A. & Salakoski T. (2018) Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies , pp. 133–142.

[40] Bader J.L. & Theofanos M.F. (2003) Searching for cancer information on the internet: analyzing natural language search queries. Journal of medical Internet research 5, p. e31.

[41] Warner J.L., Anick P., Hong P. & Xue N. (2011) Natural language processing and the oncologic history: is there a match? Journal of oncology practice 7, pp. e15–e19.

[42] Denny J.C., Choma N.N., Peterson J.F., Miller R.A., Bastarache L., Li M. & Peterson N.B. (2012) Natural language processing improves identification of colorectal cancer testing in the electronic medical record. Medical Decision Making 32, pp. 188–197.

[43] Housley W., Webb H., Williams M., Procter R., Edwards A., Jirotka M., Burnap P., Stahl B.C., Rana O. & Williams M. (2018) Interaction and transformation on social media: the case of twitter campaigns. Social Media+ Society 4, p. 2056305117750721.

[44] Jin Y. (2017) Development of word cloud generator software based on python. Procedia engineering 174, pp. 788–792.

9. APPENDIX

Table 11. Cancer keywords used for categorisation

Language	**Set of keywords**
English	Cancer, Tumor, leukemia, oncology, chemotherapy, melanoma, sarcoma, neuroblastoma, Paraganglioma, retinoblastoma, astrocytomas, retinoblastoma, lymphoma, metastasis, malignant, Telemedicine, ablation, cancer survivor
Finnish	Syöpä, kasvain, säteily, syövän selviytyjä
Swedish	Kräftan, Cancer överlevare
Norwegian	Kreft, kreft overlevende
Danish	Kræft, kræft overlevende

Table 12. Cancer types keywords used for categorisation

Language	**Set of keywords**
Stomach	Stomach cancer, Gastrointestinal, tract Cancer, colorectal, Colon cancer, stomach adenocarcinoma
Breast	Breast cancer, breast adenocarcinoma, carcinoma in situ, ductal, LCIS, nipple termed Paget, medullary
Lung	Lung cancer, squamous cell carcinomas, bronchial carcinoids
Skin	Skin cancer, lymphoma, melanoma, basal, dermatology, melanoma, moles
Blood	Blood cancer, leukemia, leucocythaemia, leucocythaemias, leucocythemia, leucocythemia, hematologic
Head and Neck	Head and neck cancer, Head and neck neoplasm, pharynx, larynx
Brain	Brain cancer, acoustic Neuroma, metastatic Brain Tumors, pituitary Tumors, oligodendroglioma, primitive Neuroectodermal, craniopharyngioma, medulloblastoma
Bone	Bone cancer, bone neoplasm, osteosarcoma, ewing tumor, fibrosarcoma, histiocytoma, chordoma
Pediatric	Pediatric cancer, pediatric cancer, childhood cancer, child cancer, wilms tumor, osteosarcoma, retinoblastoma

Table 13. Gender related keywords for categorisation

Language	**Set of keywords**
Male	Guy, spokesman, chairman, men, him, hes, his, boy, boyfriend, boyfriends, boys, brother, brothers, dad, dads, dude, father, fathers, fiance, gentleman, gentlemen, god, grandfather, grandpa, grandson, groom, he, himself, husband, husbands, king, male, man, Mr, nephew, nephews, priest, prince, son, sons, uncle, uncles, waiter
Female	spokeswoman, chairwoman, women's, actress, women, shes, her, aunt, aunts, bride, daughter, daughters, female, fiancee, girl, girlfriend, girlfriends, girls, goddess, granddaughter, grandma, grandmother, herself, ladies, lady, lady, mom, moms, mother, mothers, Mrs, ms, niece, nieces, priestess, princess, queens, she, sister, sisters, waitress, wife, wives, woman

Table 14. Gender related keywords for categorisation

Language	**Set of keywords**
13:18	My best friend, My boyfriend, My daddy, My mommy, My prom, Reaching 18, My school, My homework
19:22	semester, Reaching 20, 21st, library, My campus, My apartement, My college
23:29	at work, interview, My office, beer, My wedding, My house
30:65	My family, My kids, My son, My daughter, prayer, My daughter, My grand daughter, My grandson, for a lifetime, proud, My husband, My wife, My niece, veteran, god, pray, religion

Table 15. List of tools used in text analysis

Application	**Tools tested**	**Tools used**
Translation	Google API, Googletrans	Googletrans: `https://pypi.org/project/googletrans`
Verifying translation	Pyenchant	Pyenchant: `https://pypi.org/project/pyenchant`
Part of speech tagging	Stanford tagger, nltk taggers, GATE	Stanford tagger: `https://nlp.stanford.edu/software/tagger.shtml`
Named entity detection	Stanford ner, nltk named entity detection	Stanford ner: `https://nlp.stanford.edu/software/CRF-NER.shtml`
Tree parsing	Stanford parser, nltk tree parser	Stanford parser: `https://nlp.stanford.edu/software/lex-parser.shtml`
Hyponyms	Wordnet	Wordnet: `http://www.nltk.org/howto/wordnet.html`
Topic detection	Latent Dirichlet allocation	Latent Dirichlet allocation: `https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation`
Sentiment analysis	IBM watson Natural language understanding, Sentistrength, WNAffect (wordnet)	SentiStrength: `http://sentistrength.wlv.ac.uk`
Database	Firebase, SQLite	Firebase: `https://firebase.google.com`