Mining Health Discussions on Suomi24

Moamen Ibrahim
Centre for Machine Vision Research
University of Oulu
Oulu, Finland
moamen.ibrahim@student.oulu.fi

Matti Eteläperä
Pepron Software Services Ltd.
Oulu, Finland
matti.etelapera@gmail.com

Sercan Turkmen
Center for Machine Vision Research
University of Oulu
Oulu, Finland
sercanturkmen@outlook.com

Mina Maged
Center for Machine Vision Research
University of Oulu
Oulu, Finland
mina.ghobrial@student.oulu.fi

Mourad Oussalah
Center for Machine Vision Research
University of Oulu
Oulu, Finland
mourad.oussalah@oulu.fi

Jouko Miettunen
Lifelong Health Research Unit
University of Oulu
Oulu, Finland
jouko.miettunen@oulu.fi

***Abstract*—This paper discusses an effective way to fuse multiple tools to make a modern and easy-to-use software to get some key findings about health related topics on online social platforms, with a special focus on the famous Finnish forum, Suomi24. Several natural language processing have been tested and, sometimes, modified in order to accommodate the Finnish language linguistic structure and achieve our tasks of mining the health discussion content. We also explore the ability to monitor and track diseases in Finish open discussion forum. The developed modular system can help clinicians and medical experts to analyze similar forums to identify and track related events that can be correlated with hospital dataset in order to generate new hypotheses that initiate future treatment based approaches.**

Keywords—Natural language processing, sentiment analysis, disease detection, Sumoi24

I. INTRODUCTION

Social media is changing the nature and speed of health care interaction between individuals and health organizations. The general public, patients, and health professionals are using social media to communicate about health issues [3]. Diseases are one of the main issues people share their fight with and ask for help or any other type of support from others. Many health authorities have also started using social media platforms, including, twitter, to communicate directly to patients and interact with them [4]. In a review conducted by Stanford University regarding the use of social media and mHealth technologies for cancer prevention, cancer treatment, and survivorship, the authors highlighted clear advantages in reachability, scaled delivery and low resource setting, which enables health authorities to develop supportive social networks that connect patients and providers, encourage adherence to cancer care, and collect vast quantities of data for advancing cancer research [5].

The exploration of medical and health-related text has seen an important rise in big data applications and text analysis. This usually helps medical and clinical professionals to identify key parameters and calls for appropriate clinical decision making.

This paper aims to shed light on medical-related discussions in popular open Finnish anonymous forum, Suomi24 [2, 17]. The anonymity enables users to discuss topics and issues that may sound taboo in real life, which offers interesting perspectives for forum-based textual analysis. Especially, the paper aims to answer important questions such as identifying main health issues that worry citizens in Finland, their key worries and relieving topics. Identifying how these questions can be answered through textual analysis is by itself a challenge too that we intend to explore in detail in this paper.

Intuitively, the ability to early detect diseases or propagation trends is of paramount importance to provide efficient treatment to citizens. Therefore finding new ways to detect and track potential diseases in patients would be extremely beneficial. We hypothesize that user's behavioral change in an online forum when discussing a specific health issue provides insights into the likelihood of the occurrence of the underline disease (s) in the community. For instance, some researchers used the real-time nature of Twitter to detect events using machine learning models and send messages to those interested to receive such information [8]. Similarly, cancer patient Twitter users often share treatment experience, clinical effectiveness, financial burden, family worries, lifestyle with other patients, close friends, and relatives. This can be identified through the mining of Twitter messages. On the other hand, the growing digital record, including doctor-patient records in hospital databases has led to the proliferation of electronic health records (HERs)

Mining EHR has the potential for establishing new patient stratification principles and for revealing unknown disease correlations. However, a broad range of ethical, legal and technical reasons currently hinder the systematic deposition of such a dataset [6].

Textual analysis of forum data or HER calls for natural language processing (NLP)-like analysis, which is a subfield of artificial intelligence, that enables the machine to comprehend the meaning of textual input. Nevertheless, when it comes to the Finnish language, many NLP developed tools are still lacking efficiency as compared to English NLTK tools [9]. Two potential approaches can be distinguished for this purpose, either translate the Finnish text into English and use efficient NLP tools developed therein, despite the limitation of such automatic translation (e.g., GoogleTrans API [11]) or comply to existing Finnish parser tools and acknowledge the inherent limitations. This includes, for instance, FinnPos for lemmatization and morphological tagging, Turku neural dep parser, AFINN for sentiment analysis.

II. Dataset handling

To analyze suomi24, we used the online corpus provided by the University of Helsinki, which had all discussions from 2001 till June 2015 with around 231 million sentences. As our focus is on health-related topic online, we looked for 'Terveys' subtopic which means health in Finnish, compared to other approaches which are catching topics from the forum using word matching with a list of health-related keywords, subtopic extraction enabled us to get a relatively big data-set of specifically health discussions with less noise in the data-set. The final data-set was around 28 thousand sentences to be used for actual analysis. To make it easier for our future tasks, we developed an architecture to help us progress in such a project. The architecture discussed in Figure 1, explains how codes and tools will be used in the subsequent steps to achieve the purpose of our analysis. Such architecture will be helpful for future modifications and will provide a good structure for our codes. The architecture divides the tools to Finnish and English, which makes it easier to use, edit or adjust later on. For example, in the English section, we added the translator, LDA topic detection and Stanford tools. This approach helped the collective work in the team. Also, for the Finnish tools, we added the AFINN [16], Polyglot [12], Turku's parser tools [15] and FinnPos. The graphical user interface (GUI) was an independent system to communicate with our wrapper (main.py) which runs the tools and scripts. This also helped, for instance, to maintain and monitor the initial set of keywords in both Finnish and English. Of course, we needed such a system, helpers like keywords in Finnish or English or lists of significant words. We believe that the current architecture was helpful and saved a lot of time and effort for us. The data in Suomi24 text dumps was supplied in JSON format and the storage size of the raw data was about 15 gigabytes.

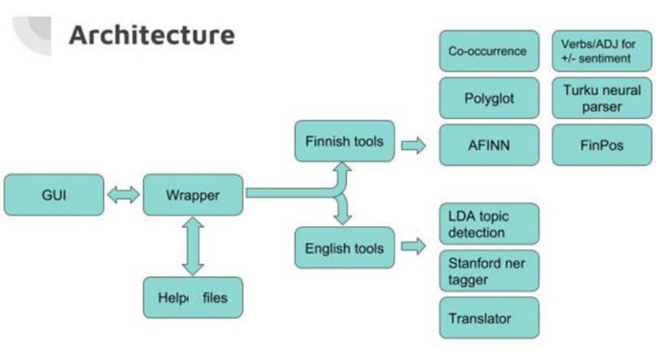

Fig. 1. General architecture for data collection and analysis

An example of dataset structure is shown in Fig. 2


```
{
    "body":"<p>Niin k\u00e4ynki ens kerralla! Niin kannattaa
            teh\u00e4 muittenki. Tai hakea K\u00e4rkk\u00e4iselt
            \u00e4.</p>",

    "quote_id":65033569,
    "deleted":false,
    "created_at":1388253658000,
    "comment_id":65288763,
    "anonnick":"kuparia kuparia",
    "thread_id":11858119,
    "parent_comment_id":65032642
}
```


Fig. 2. Suomi24 message data example.

Preprocessing

First for the purpose of subsequent analysis, we divided the dataset on monthly basis in order to identify monthly change of patterns. More specifically, the following were considered for preprocessing task.

i) Filtering out the data threads only to take "Terveys" (eg. "Health") topics into account.
ii) Conversion of JSON message body data into new line divided raw text files, named month-year.txt.
iii) All metadata was stripped.
iv) Removal of HTML tags from body text
v) Conversion of text such that each word on a new line
vi) Using the FinnPos ftb-label tagger, we lemmatized the new-line separated texts.
vii) Using cut and sed tools on Linux bash, we stitched the lemmatized newline separated words back to sentences.

III Named-entity and co-occurrence analysis

To comprehend the content of the collected health dataset, the first step was to inquiry about the named-entities conveyed in the dataset. Intuitively, the existence of high frequent named-entities would indicate some dominance of user-discussion towards topics involving such named-entities. We, therefore, used Stanford named-entity tagger [14] to identify the various types of entities (e.g., person, location and organization). We denote by PER, LOC, and ORG the named-entities corresponding to person, location and organization, respectively.

The results shown in Figure 3 indicate a strong dominance of location (LOC) entities which constitute almost 77% of the most frequent thirty named-entities of the corpus. However, the most frequent person-entities are "Soini" and "Johan", which can, therefore, be employed to support any potential hypothesis through a statistical-based reasoning.

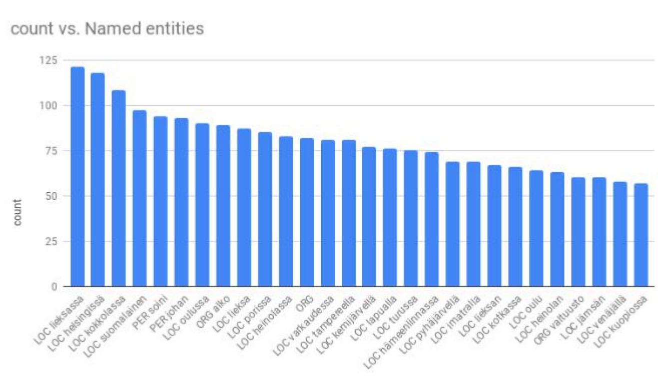

Fig. 3. Frequency of 30 highest named entities.

Next, we would like to inquiry about the actions induced by these named-entities. For this purpose, we employ parser-tree to identify the main verbal expressions associated with these named-entities. Using Turku neural parser, we were able to visualize the main sentences associated with these named-entities and identify common verbs and adjectives associated with them. This is exhibited in Figure 4.

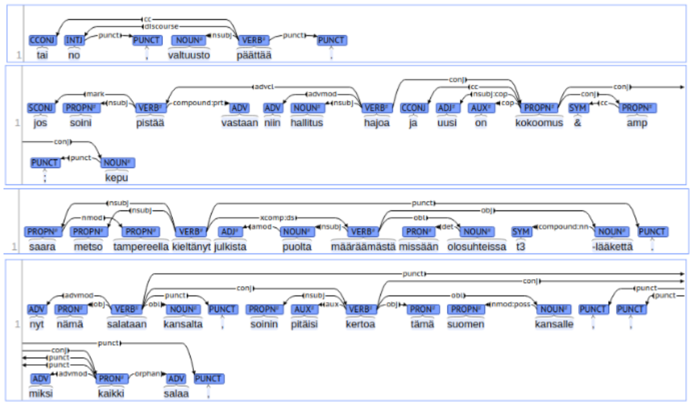

Fig. 4. Instances of Parser tree on 30 most common named entities.

The approach will, therefore, provide a user-friendly interface for the decision-maker to visualize the outcome of this parser to comprehend the actions associated with most frequent named-entities.

Another trend in comprehending the context of the discussion is to investigate the frequency of the co-occurring words, which can also provide useful insights about the salient and dominant topics of the discussion forum. For this purpose, using a simple word-count based approach that accounts for words situated next to each other, the overall word-pair count is depicted in Figure 5. From an implementation perspective, to perform this task, we made a script that assumes any new given string is a paragraph, tokenizing the paragraph into sentences. After that, it tokenizes the sentences into words, and removes any odd characters using regex-functions, as well as 'stopwords' (currently declared as a list of words), and, finally reconstructing the sentences, accordingly. Next, we loop over the sentence twice and add co-occurrences of every word pair into a dictionary. Therefore, Figure 3 shows the most common co-occurrences in the health-related topics of Suomi24.

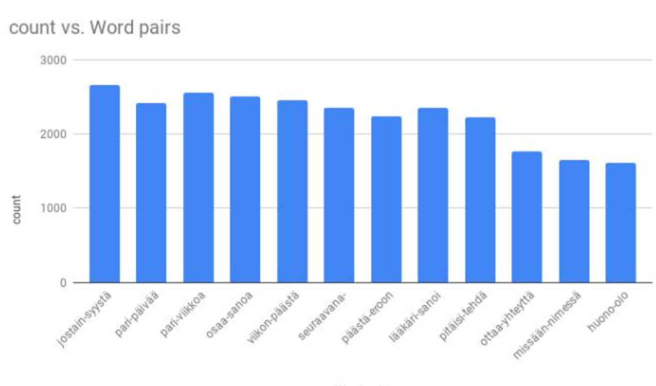

Fig. 5. Count vs word-pairs

The analysis of the data count of Figure 3 reveals the dominance of the phrase constructs (for the first half of pairs in Fig 3), but also for food patterns (pästää-eroon), advice-type recommendations (pitäisi-tehdä, otta-yhteytää) and stress on doctor's recommendation (lääkari-sanoi). The preceding highlights the importance of food, social support and clinical recommendation in the overall discussion forum. Therefore, any approach to patients through networking should ideally take into account such a pattern.

IV Sentiment Score based Analysis

Sentiment analysis has been conducted both at the global level and monthly level to help the analyst to focus on a specific pattern of interest.

For this purpose, we use a modified version of AFINN [13]. The result of the global trend in sentiment across all collected dataset is shown in Figure 6. Typically, one computes the sentiment score for each textual chunk of the dataset, and we report the proportion of sentiment score whose value takes a specific score. The results vary from -35 to +30, with more tendency to negative emotions as shown in figure 6. More than 40% of the sentiment from sentences was in the range between -10 and zero.

From an implementation perspective, the following steps were taken to implement this task. We retrieve a list of a lemmatized words of positive and negative sentiment verbs as well as adjectives. For this purpose, initially, a translated word list (from English) was used, but later on, another source file for positive and negative sentiment can be downloaded from https://www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages/home Sentiment analysis.

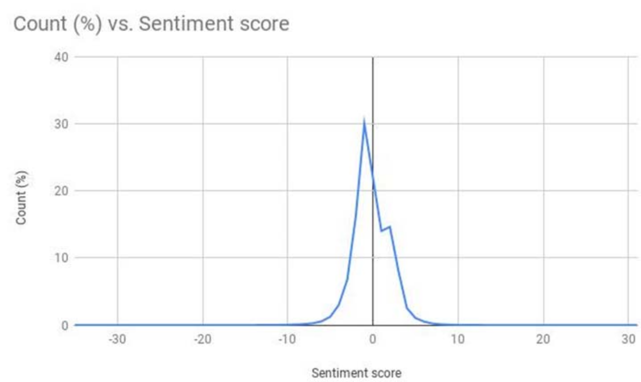

Figure 6. Overall sentiment score

While the monthly averaged score of the sentiment is reported in Figure 7. In this case, we were looking for variation in the average sentiment per month from the year 2001 to 2015. From reading the results of Fig 7, one notices a big spike in the positive sentiment on the beginning of the year 2006, then sentiment dropped to -0.3 in the year 2007. The biggest drop in sentiment was in the year 2009 with significantly less than -0.37 in the sentiment average score, followed after with another drop by middle of the year 2010. These changes have also found to correlate with the economy downturn and the introduction of new changes in the Finnish welfare system that trivially affected citizens' sentiment.

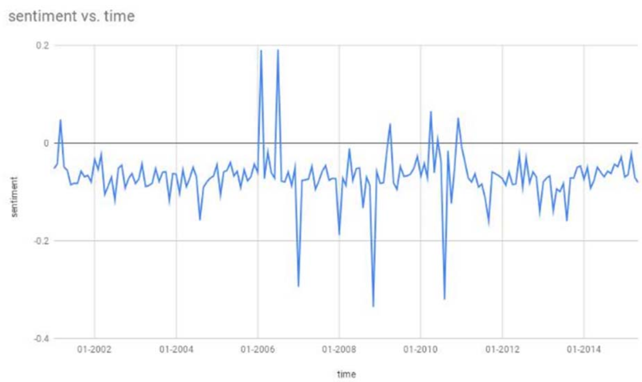

Figure 7. Monthly variation of sentiment score

V Topical Analysis

In this step, our focus is to identify the topics and their variation with time. For this purpose, we performed LDA [10] directly on Finnish text as we constructed the corpus using every sentence in every month and from which we were able to train the Genism model with such words. Results varied from one month to another due to the variation in the number of threads related to health. This was one of the biggest challenges for LDA topic detection using this approach. The approach also makes use of lemmatization and stopword removal. The results shown in Table 1 provide useful insight into the content of the discussion.

Table 1. Topic detection using LDA after lemmatization.

Year	Topics
04-2009	pysäköintialue, pysähtyä, pysähtytyä, pysähtyneisyys, pysähtyneen, pysähdys, pysyä, pysyvät, pysyvästi, pysyvä, pysytellä, pysyntyä, pysyikö, pysyi, pysy, pystyä, pystyyn, pysäköintihalli, pysäkki, pysäköinti, voida, quo, mennä, tehdä, tietää, käydä, ihminen, käyttää, päivä, ottaa, lääkäri, sanoa, haluta, tuntua, aiheuttaa, syödä, päästä, lääke, elämä
05-2009	putki, putkahtaa, putkaan, koska, putka, putiikki, pussillinen, pussiin, pussi, puskuri, puskea, puska, push, pururata, pursuta, pursuilla, pursi, purra, purkutuomio, puuhataan, quo, voida, tehdä, mennä, käydä, lääkäri, ihminen, tietää, päivä, ottaa, sanoa, syödä, lääke, haluta, päästä, käyttää, viikko, auttaa, paikka
11-2013	pystyis, pyssymestari, pyssy, pyssätä, pyssyä, pystynyt, pystyisi, pystyisikä, pystyisikään, pystyisit, ei, pystyitää, pystykahvi, pystykkä, pystykuolema, pystykä, pystymätön, pystyneet, pystyidä, pyssystään, jooh, pyssyyn, suanoo, quo, voida, tehdä, mennä, ihminen, käydä, ottaa, tietää, sanoa, haluta, päivä, työ, mies, päästä, maksaa, raha, kaupunki, käyttää, paikka
12-2013	pääministeri, päältä, päältäni, pääminisiteriksi, päänahassani, pääministerimme, pääministerin, pääministerinä, päämäärä, päämäärän, päämääränä, päämäärättömästi, päämäärää, päämääräänsä, pään, päänahan, päämiehen, päältään, ja, ei, että, kun, se, niin, quot, mutta, ole, jos, oli, nyt, en, tai, kuin, olen, voi, sitten, olla
11-2014	pyyhkainen, pyyhkeeksikään, pyyhkidä, pyyhkiessä, pyyhkii, pyyhkiin, pyyhkijä, pyyhkijää, pyyhkiminen, pyyhkityne, pyyhkiytyä, pyyhkiä, pyyhkäa, pyyhkäisnätä, pyyhkäistä, pyy-hältävä, pyyhältää, pyyhitty, pyyhkee, pyykinpesuaine, quo, voida, tehdä, mennä, ihminen, käydä, ottaa, tietää, työ, haluta, sanoa, mies, n, kaupunki, elämä, suomi, päivä, maksaa, raha, päästä, lähteä, jäädä, käyttää
12-2014	pyydyksiä, pyyhkiminen, pyydys, pysäyttää, pyydetään, pyy-detty, pyydellä, pyyde, pyy, pytä, pytytyä, pytyn, pyttyy, pyt-tymäinen, pytty, pyttis, pyttipannu, pyydystäjä, pyyhkimään, pyyhkii, quo, voida, tehdä, mennä, ihminen, käydä, ottaa, tietää, haluta, sanoa, kaupunki, työ, päivä, mies, raha, suomi, maksaa, kunta

For instance, topics like "doctor", "medicine", "life", "exercise", "money", "city", "goals", "municipality", "work" are showed up significantly as a result of LDA based analysis.

VI Disease monitoring and tracking

To track the disease that can occur in Suomi24 collected dataset, the following steps have been carried out.

i) Generate disease vocabulary in Finnish
ii) Match disease vocabulary to lemmatized Suomi24 data per month
iii) Generate year and total occurrences
iv) Normalise the number of monthly and yearly occurrences of diseases with the number of messages for each month and year

All the above operations were performed on the Linux command line with sed, cut, etc. tools. Disease vocabulary was obtained from the FinnMesh Medical Subject Headings ontology. The ontology also includes a lot of other health-related words. Overall the vocabulary was 80k words and it was generated by matching “prefLabel” and “altLabel” tags of the mesh-skos.rdf file.

Matching this amount of words to monthly Suomi24 data proved to be too time-consuming, so filtering of the vocabulary was needed. This operation involves the following:

- Match the 80k word disease vocabulary to one month of Suomi24 data. This step took about 20 minutes.
- Count the number of occurrences per word. Remove words that have 0-2 occurrences.

As a result, the vocabulary was reduced from 80k to a much more manageable 1722 words. Even in this list, the number of words directly related to diseases was small. Besides, the list was manually checked and 143 disease names were identified.

Figure 8 shows the yearly mentionings of different diseases. In the graph, it can be seen how the amount of data in Suomi24 expanded around the year 2011. The most mentioned disease in the data is "masennus", e.g., depression. Subtypes of depression such as "vakava masennus" e.g., severe depression were grouped to the data. The same was done for other disease super-classes also, and these are shown in the graph in capital letters.

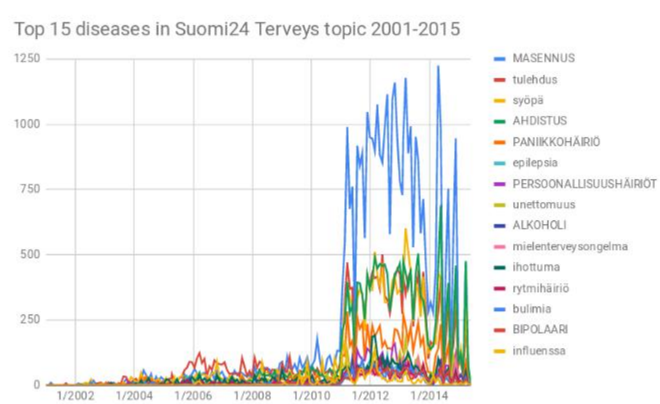

Fig. 8. Top 15 diseases in Suomi24 Terveys topic 2001-2015.

Figure 9 depicts the most mentioned disease per month. This data is not normalized with respect to the amount of text generated each month.

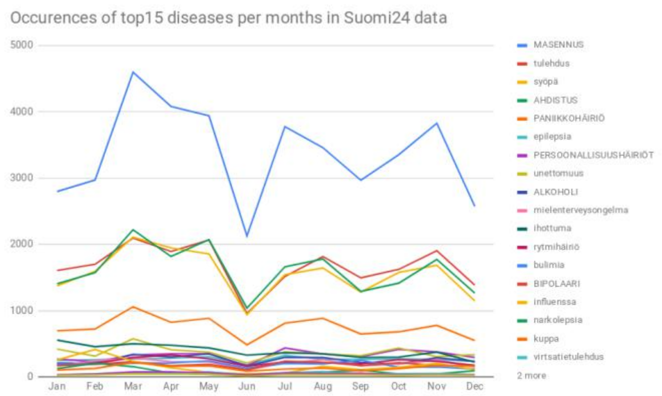

Fig. 9. Occurrences of top 15 diseases per month in Suomi24 data.

The main point was to identify the most common diseases. For this purpose, we used the same ontology to detect the repetition of diseases among our data-set, we were then able to identify five mainly common diseases as "masennus" ("depression"), "syöpä" ("cancer"), "ahdistus" ("anxiety"), "paniikkihäiriö" ("panic disorder") and finally, "epilepsia" ("epilepsy"). This shows that that three out five common diseases in Suomi24 were related to psychological behaviors. This finding is also in line with recent reports of World Health Organization [18] that highlights the dominance of psychological and mental related health issues worldwide.

Furthermore, the application of parser tree on some sample sentences that have most common disease terms gave us indication about the related words to these disease terms. This can also indicate about key worries associated to specific diseases. As shown in Figure 8 and 9, we were able to extract some main worries that concern people on online forums, like "masennus" as well, which makes it both a disease and a key worry associated to other diseases like cancer.

VIII. CONCLUSION

In this paper, 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 understanding and findings about health related topics raised on Suomi24, the most popular Finnish anonymous online forum. Given the absence of effective NLP tools for Finnish language, a new framework has been put forward and implemented in order to mine health topics in Suomi24. The limitation of the developed tool has also been commented. Our analysis identified the most important diseases, health related topics and comprehended the occurrence of such diseases in the corpus by analyzing co-occurring terms. Temporal analysis has revealed significant changes either in sentiment analysis, disease detection and common topics. Therefore, fusing all these data would provide a more complete picture about the health topic occurrences in the online forum[1].

ACKNOWLEDGMENT

This work is partly supported by Finnish Cancer Foundation on Psychosocial factors on cancer community (2017-2019), and EU YoungRes (#823701)

REFERENCES

[1] S. Finland, “Use of information and communications technology by individuals,” Helsinki: Statistics Finland. Retrieved June, vol. 8, p. 2016, 2015.

[2] I. Khaldarova, S.-M. Laaksonen, and J. Matikainen, “of the publication: Type of the publication,” series: Media and Communication Studies Research Reports 3/2012, 2012.

[3] R. Thackeray, B. L. Neiger, C. L. Hanson, and J. F. McKenzie, “Enhancing promotional strategies within social marketing programs: use of web 2.0 social media,” Health promotion practice, vol. 9, no. 4, 338–343, 2008.

[4] C. L. Ventola, “Social media and health care professionals: benefits, risks, and best practices,” Pharmacy and Therapeutics, vol. 39, no. 7, 491-, 2014.

[5] J. J. Prochaska, S. S. Coughlin, and E. J. Lyons, “Social media and mobile technology for cancer prevention and treatment,” American Society of Clinical Oncology Educational Book, vol. 37, pp. 128–137, 2017.

[6] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care,” Nature Reviews Genetics, vol. 13, no. 6, p. 395, 2012.

[7] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, “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, 2010, pp. 841–842.

[8] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes twitter users: real-time event detection by social sensors,” in Proceedings of the 19th international conference on World Wide Web. ACM, 2010, pp. 851–860.

[9] S. Bird and E. Loper, “Nltk: the natural language toolkit,” in Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004, p. 31.

[10] X. Qiu and C. Stewart, “Topic words analysis based on lda model,” arXiv preprint arXiv:1405.3726, 2014.

[11] M. Aiken and S. Balan, “An analysis of google translate accuracy,” Translation journal, vol. 16, no. 2, pp. 1–3, 2011.

[12] R. Al-Rfou, B. Perozzi, and S. Skiena, “Polyglot: Distributed word representations for multilingual nlp,” arXiv preprint arXiv:1307.1662, 2013.

[11] Github link for the souce code is available at https://github.com/moamenibrahim/nlp-project

[13] R. Al-Rfou, V. Kulkarni, B. Perozzi, and S. Skiena, “Polyglot-ner: Massive multilingual named entity recognition,” in Proceedings of the 2015 SIAM International Conference on Data Mining. SIAM, 2015, pp. 586–594.

[14] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. Mc-Closky, “The Stanford core nlp natural language processing toolkit,” in Proceedings of 52nd annual meeting of the association for computa-tional linguistics: system demonstrations, 2014, pp. 55–60.

[15] J. Kanerva, F. Ginter, N. Miekka, A. Leino, and T. Salakoski, “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, 2018.

[16] SILFVERBERG, M., RUOKOLAINEN, T., LINDÉN, K. and KURIMO, M., 2016. FinnPos: an open-source morphological tagging and lemmatization toolkit for Finnish. Language Resources and Evaluation, 50(4), pp. 863-878.

[17] Lagus, Krista, Pantzar, Mika, Ruckenstein, Minna & Ylisiurua, Marjoriikka (2016) ’Suomi24 – muodonantoa aineistolle’. Helsinki: Helsingin yliopiston valtiotieteellisen tiedekunnan julkaisuja 10.

[18] World Health Organization, Mental Health Atlas, 2018, available online at https://apps.who.int/iris/bitstream/handle/10665/272735/9789241514019-eng.pdf