Mining Security discussions in Suomi24

Eetu Haapamäki
Faculty of Information Technology,
University of Oulu,
Oulu, Finland
eetu.haapamaki@gmail.com

Juho Mikkola
Faculty of Information Technology, University of Oulu
Oulu, Finland
juho.mikkola@fewiz.com

Markus Hirsimäki
Faculty of Information Technology, University of Oulu
Oulu, Finland,
hirsimaki.markus@gmail.com

Mourad Oussalah,
CMVS, Faculty of Information Techology, University of Oulu
Oulu, Finland
mourad.oussalah@oulu.f

***Abstract*—This study examines how social network based approach can be applied in order to mine the security oriented discussions in Suomi24 online forum. The approach employs a student survey questionnaire to collect a dictionary related to Finland national security. In subsequent analysis, the vocabulary terms are mapped to Suomi24 corpus in order to construct the associated social network analysis that quantifies the dependency among the various vocabulary terms. Especially, the analysis of the dynamic variation of the network topology would enable the decision-maker to devise appropriate communication scheme to maximize intervention in the public sphere and reach a wider audience. Besides, a parser that finds the keywords from VeRticalzed text data format is developed to aid the construction of the underlined social network.**

Keywords—Suomi24, Security threads, Social Network

I. INTRODUCTION

Since the introduction of web 2.0, the user-generated content has become increasing popular in the web and revolutionized the way in which the users interact with one another, yielding large number of blogs, online communities and social platforms where users share thoughts, debate, seek advices and sustain common interests, among others. Many of these networks serve as the de-facto internet portal for millions of users. According to Laaksonen *et al.* [1] internet is not only a practical data gathering tool but it is increasingly becoming a stage where people form and live culture. This, in turn, due to the large amount of data that can be collected about single user / group, offers golden opportunity to analyze human behavior on a large scale and track useful events [2]. On the other hand, such networks have also offered opportunity to malicious users and organized crimes to win support for their causes and publish illegal or suspicious content that may affect national security of countries or institutions.

Suomi24 is Finland's largest topic-centric social media, spanning 16 years (since 1998) and 6.5 millions of threads [3]. It is known as a forum for manifesting social and political frustration, critique, as well as sharing health or well-being issues, discussion on hobbies, consumer experiences and everyday life. Due to its anonymity, the users of the site participate in discussion threads without sharing their individual identities, which probably enables users to share their intimate thoughts and experiences. Indeed, Suomi24 is regarded as the most visible Finnish language internet forum that provides a public platform to discuss anything under the sun without any restriction. The latter is also motivated by the fact that in contrast to other public forums and news blogs where racist and insulting talk is prohibited, such rules are poorly enforced in Suomi24, so that "anything can be said about anybody" according to Media commentator YLE journalist [4]. The bundles of threads around various topics open opportunities for various methodological development characterizing posts, threads, and bundles using various linguistic, temporal and statistical like features. Dataset of Suomi24 (www.sumi24.fi) is owned by a Finnish publishing private company Aller Media Oy. www.suomi24.fi is the sixth most visited Finnish website with a monthly reach of circa 2.3 million users measured on all different devices in May 2018 [5]. It is the most popular Finnish internet discussion forum by website reach. The dataset is publicly available in original and lemmatized form from https://www.kielipankki.fi/corpora/. The availability and rich structure of the text has motivated the use of text analytics in order to mine the discussions and provide insight on individual and community behavior, discover new patterns, test new hypotheses, marketing activities and personalized content creation. For instance, the website has been explored to identify health patterns in Finnish society [6], business-behavior [7], among others.

In this paper, we are mainly interested into exploring Suomi24 dataset for security related discussions in a way to provide some insights that enable decision-makers, security officers to comprehend community interests and worries in this matter as well as identifying some potentially hidden patterns. Strictly speaking, there are several reasons that prompt the interest of security related discussion in such forum. First, security events like Paris bombing or refugees crisis intuitively spark public interest at wide, which is then materialized in their messages in blogs/forums. Second, discussion forums seem like an ideal place to study how people conceptualize security issues because people perceive them private enough to express thoughts that they might not say in a more formal face-to-face research interview [8]. Third, Finland is perceived as one of the safest countries in Europe whereas the level of preparedness of Finnish welfare state for refugees and terrorism trials is a sensitive issues that is well documented in national political sphere, and subsequently in public debate as well. This opens up the debate whether Finland is a safe country for its citizens. Fourth, according to European Commission's Eurobarometer

Survey [9], Finns perceive the risk of a terror attack lower than Europeans in general (11% versus 40%,), which makes any security event subject to possibly biased and non-rational debate. Fifth, for historical reasons, the perception of Russian threat in both public and political compass is very sensitive, and accordingly well commented in open forums. The key research question is

How do (Finnish) citizens conceptualize security issues in Suomi24 internet discussions?

For this purpose, a methodology that combines keywords based analysis and social network like analysis is devised and tested. For instance, finding out what security related topics are being discussed and how often they are discussed can provide useful insights on the strength of such (sub) topics.

II. METHOD

A. Data

The data consists of a sample of Suomi24 posts [3] between 2001-2014. There are 2 127 506 forum posts with 10 112 531 sentences in total.

The format of the data is VeRticalized Text (VRT), which consists of XML-style entities which contain a format that expresses the text by making every single word or punctuation in a new line in followed by related information to the text entity separated by tabs. The structure of the format is expressed in Figure 1. The data consists of 1 955 separate VRT-files, with approximately 800-1800 separate posts per file.

```
<text>
  <paragraph>
        <sentence>
        This
        is
        a
        sentence
        .
        </sentence>
        <sentence>
```

Fig 1. VRT data structure

B. Method

1) Keyword Identification

The first step in our methodology is to identify the key-words related to national security. For this purpose, a survey-based analysis was performed. More specifically, student population were asked using both mailing list and online-platform the question "What words and concepts you think are related to national security of Finland?"

Each respondent can input as much as he wants of individual keywords that he believes relevant to his conceptualization of National security of Finland. Most students have inputted more than 100 words. Besides, no constraints have been imposed to the students, enabling any forms of interaction among the students to discuss such keywords if any.

The full list of key-words that are shared by more than 30% of the respondents are reported in the appendix of this paper, which contains a total of 159 distinct keywords (See Appendix). A simple scrutiny of this list reveals the following:

- The students encompass a much wider conceptualization of the notion of national security that goes beyond a strict lexical refinement of arms, conflict and terror.
- Finland's neighbor countries and even national organizations have been listed to be linked to Finland national security. This includes Russia, St Petersburg, Sweden and even Europe and Estonia, Finnish parliament, EU institutions, Vladimir Poutine, among others.
- Surprisingly, cultural identities like Islam, Feminism, communism were also seen as linked to national security concerns.

Figure 2 below summarizes the main classes that can be assigned to the keyword list.

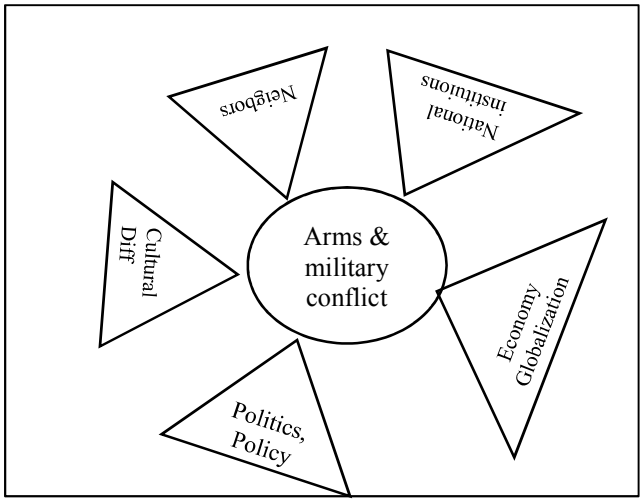

Fig 2. Categorization of national security keywords issued from students' survey

2) Social Network-based analysis

Once the list of keywords is established, it is inserted into the parser, which tags the associated posts in Suomi24 with the matching keywords and collects them into threads. The next phase in our methodology is to analyze the interaction and co-occurrence of these keyword terms in Suomi24 corpus. For this purpose, a social-network based approach is employed. The essence of such approach is to assume that each keyword as a node and whenever two keywords are mentioned together in the same thread, an edge between them is established. This creates a social network highlighting the interacting among these keywords. The rationale behind such reasoning lies in the rich and wide spread of the collected keywords that encompass most relationship types that may occur in standard human

discussions. For instance, if "terror" is mentioned alone in the post, its impact would be less if it is mentioned in connection with Finnish parliament, EU institution, or even a place (i.e., Finland, Helsinki, neighbor countries), all of which are also part of the keyword list. Therefore the "edge" link posits as a supporting evidence, according to some relationship type, that supports the given keyword. Accordingly, if a given keyword is found in some thread but has no support in terms of other keyword across any of the Suomi24 threads, will be automatically removed from the graph representation.

On the other hand, the topology of the network can also be made dependent on the number of posts where the two keywords co-occur; that is, an edge between two keywords is established only if the number of posts containing the two keywords is beyond a certain threshold T. Therefore, an iterative approach is established in order to track the network properties where the threshold is set initially to one and incrementally increases until no edges were established. At a given network configuration, we report a set of metrics that characterize the underlined network. This consists of: number of nodes, number of edges, maximum degree, average degree, global clustering coefficient, diameter, average path length, size of giant component, size and number of communities and the associated quality measure [10]. The process is illustrated in Figure 2.

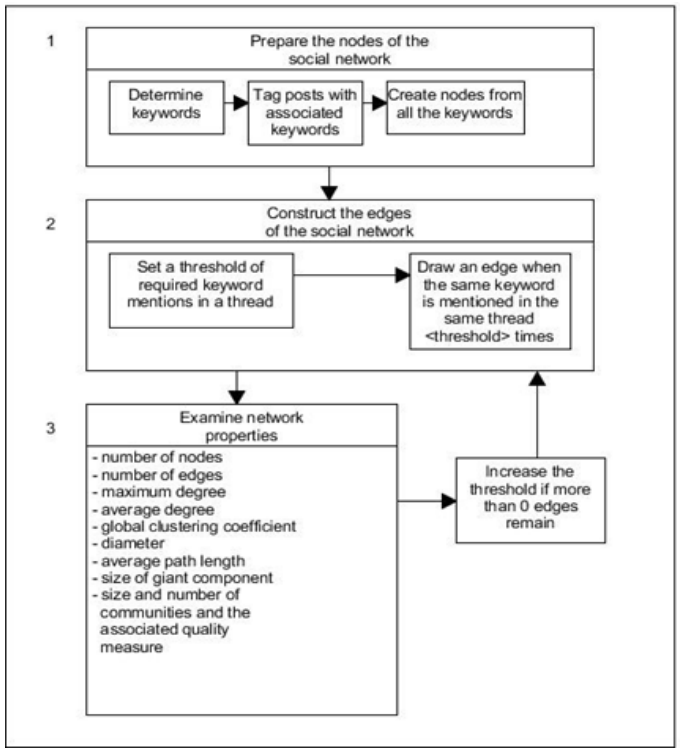

Fig 3. Process illustrated

III. IMPLEMENTATION

In order to process the VRT format of Suomi24 posts, we implemented our own VRT parser that fit the needs of the project. The parser is in python and uses the Networkx and Polyglot-packages among some common packages. The source code is available at https://github.com/Eedvard/SNA

The tool parses through the VRT-files, finds the posts that have any keywords in them, tags the posts with the keywords that are found in them and draws the social network with nodes and edge. The high level description of the workings of the code is outlined in pseudo code of Figure 4.

Parsing and creating a social network from keywords

1. Load VRT-file to memory
2. Iterate the file row by row and any XML-style tags to establish the elements that wrap the words in a post
3. Extract the wordings of the post and compare them against the defined keyword list
4. Tag the posts with the keywords
5. Create the nodes of the network from the keywords
6. Establish an edge between two nodes of the network if the number of co-occurrence of the keywords in the same thread exceed the threshold
7. Report the properties of the graph
8. Draw a graph that evaluates the evolution of these properties,

Fig 4. Description of the code for generating a social network of keywords

Next, we relied mainly on NetworkX implementation to compute the various network graph properties. The global clustering coefficient is computed using the average clustering coefficients of individual nodes of the graph. Community detection is performed using Girvan-Neuman method where the coverage of the partition is used as a metric to quantify the quality of the partition [11].

IV. RESULTS

Starting with the whole keyword list and trying to match the occurrence of these keywords in Suomi24 posts according the methodology pointed out in the previous section, the overall structure of the network graph at the beginning (threshold T=1) is illustrated in Table 1. Next, the process is iterated with various threshold levels until there is no node left in the network. The graph topology for threshold values T= 1, 300 and 800 is highlighted in Figure 5, 6 and 7, respectively.

Table 1: Initial keyword network properties

Graph Property	**Value**
number of nodes	124
number of edges	658
maximum degree	43 (‘politiikka’)
average degree	10.61
global clustering coefficient	0.32
(giant) diameter	3
(giant) average path length	1.43
size of giant component	46 nodes and 658 edges
number of communities	35
community node coverage quality measure	0.36
community edge coverage quality measure	8.2

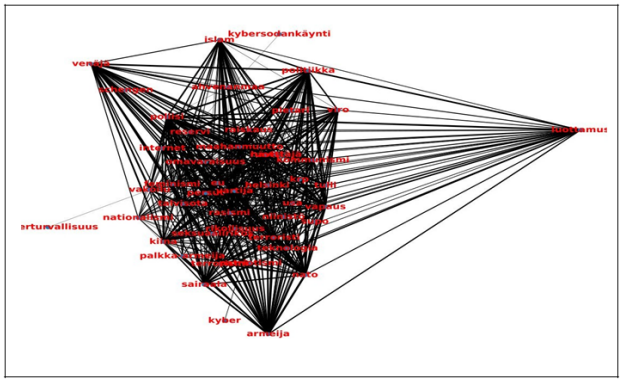

Fig. 5. Social network for Threshold =1

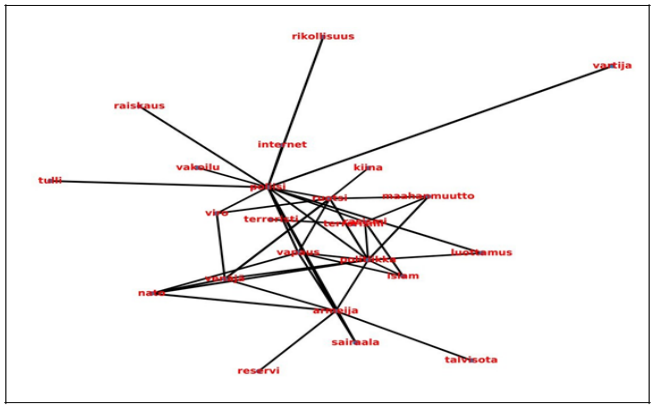

Fig. 6. Social network for Threshold T = 300

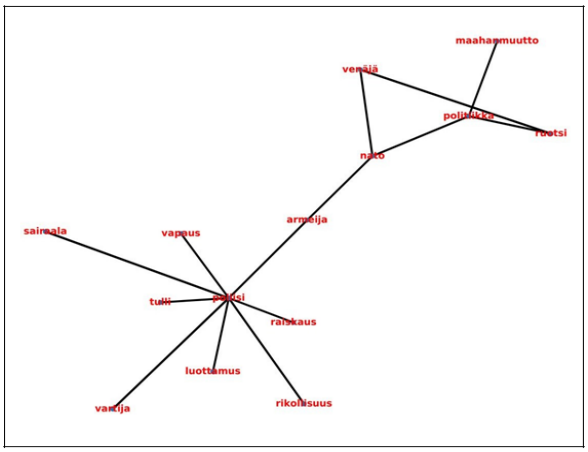

Fig. 7. Social network for threshold T=800

In the above representations, the width of the edges is expressed by the scale *log(log(x+1))* to shrink the largest weights considerably to make the graph more readable.

Table 2 summarizes the graph properties in terms of the associated metrics for specific values of threshold (1,50, 300, 1764). The latter value corresponds to the case where the network shrinks to two node graph.

Table 2. Social network properties at various thresholds

Property (Threshold T)	**1**	**50**	**300**	**1764**
number of nodes	46	34	24	2
global clustering coefficient	0.85	0.66	0.31	0
diameter	3	3	4	1
average path length	1.43	1.72	2.12	1
number of nodes in giant component	46	34	22	2
number of edges in giant component	614	183	40	1
number of edges	614	183	41	1
maximum degree	41	29	14	1
average degree	26.70	10.76	3.41	1
number of communities	35	43	14	0
community node coverage	0.98	0.89	0.54	0
community edge coverage	8.8	4.41	1.32	0

The preceding provides only a global evaluation of the various graph properties at specific threshold values. In order

to visualize the continuous evolution of the various graph property metrics, Figure 8 and Figure 9 illustrate the evolution of the various graph metrics with respect to the threshold values.

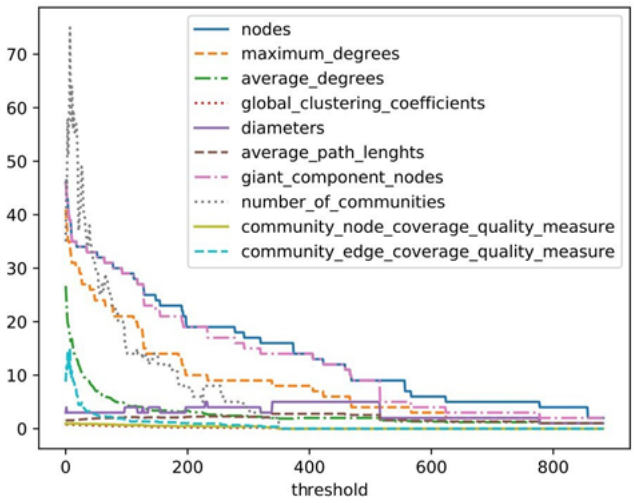

Fig 8. Evolution of network properties with respect to threshold T.

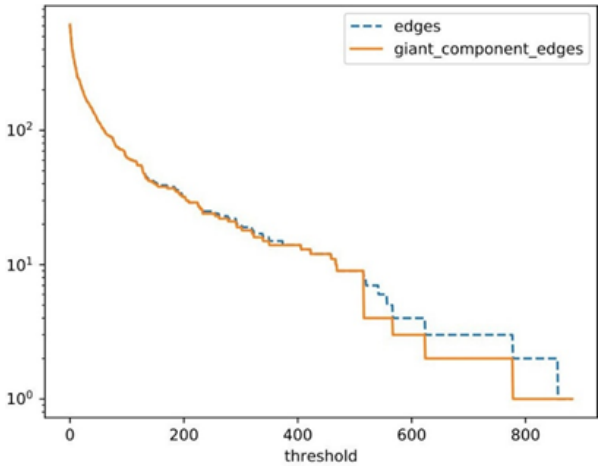

Fig 9. Evolution of graph number of edges and that of associated giant component.

Discussions

The results illustrated previously reveal the following:

i) Comparing the number of nodes of the initial graph (124 nodes) to the initial number of keywords (159) indicates that there are 35 keywords that are not involved in any coalition with the other keywords, which are therefore excluded. This occurs because either such keyword does not have any occurrence in Suomi24 posts or has no co-occurrence with any other keyword of the list.

ii) Examination of the nodes in figures 6-8 provides insights into the main wordings that have large impact in users' discussion. The key clusters of the graph are constituted of keywords: luottamus (trust), Islam, Politiika (Politics), venäjä (Russia), nationalismi (nationalism), sairaala (hospital), Armeija (army). While the strongest tie occurs between keyword polisi (police) and sairaala (hospital), which somehow reflects the key-players in tracking potential fatalities (police force) and treatment of fatalities / injuries (hospital).

iii) The preceding can be used by the decision-maker to enforce their communication strategy in order to ensure a wider audience. For instance, the importance of police and health centres in shaping their communication strategies in dealing with terror or security related event. Similar reasoning applies to Finnish political sphere as well as the sensitive neighbor Russia.

iv) Graphs in figures 8 and 9 as well as results of Table 2 indicate the smoothly decreasing property of the number of nodes of network, average degrees, clustering coefficient, average path length and community based metrics as threshold T increases. This is widely expected due to the structure of the network that constantly shrinks. Nevertheless, it is also worth mentioning that the diameter of the network changes moderately as the network shrinks as a result of increasing the threshold level.

v) Although not performed in this paper, the use of threshold value can be combined with other graph statistical measures in order to convey a dependency graph constraining the keyword list, which can then be employed for subsequent information retrieval tasks [12-13].

vi) The variation of the community level indicator indicates the quality of the community that can be extracted from the network reduces as the threshold increases. This means that if one would like to extract meaningful subnetworks from the graph, then one has to do so at the initial stage.

vii) The results obtained in this investigation in terms of the most influencing topics / keywords that best describe the security concerns in Suomi24 agree to a large extent to other alternative resulted highlighted in communication and security science. Indeed, the importance of the human element in achieving security and trust is well acknowledged [14-15]. Similarly, the research into socio-technical security and trust fields gain momentum and highlight the importance of choosing the right target group through appropriate communication terminology in order to gain public trust and strengthen societal security [17-18].

V. CONCLUSION

This paper outlines a social network based approach to map the national security related discussions occurring in the Suomi24 online forum. The developed approach relies on two pillars. First, a Finnish national security vocabulary is constructed using student online survey. Second, a social network where the nodes correspond to vocabulary terms (keywords) is constructed by matching the terms to Suomi24 corpus. Especially, whenever two keywords co-occur in the same Suomi24 thread, an edge between these two entities is established. Therefore, the dynamic analysis of the topology of the network, when introducing a confidence factor corresponding to the number of co-occurrences required for the connection to be held, allows us to devise the influencing concepts that should be taken in order to develop appropriate communication strategy with regard to national security. The findings in terms of influencing concepts have also been found to agree with results in security science and risk analysis.

In closing, we have constructed a tool that can be used to examine the strength of the connection between multiple keywords in conversations of Suomi24. The same process can be conducted using any keywords to examine a different topic as well as a larger dataset to gain different insights from the discussions in Suomi24. Raising the threshold for the required weight of the edge is useful to find out the properties of the network when the least relevant keywords are removed iteratively. These results can be then further examined by extracting the named entities to examine the locations, organizations and persons that appear in the data.

ACKNOWLEDGMENT

This work is (partially) funded by the European Commission grants) YoungRes (823701) and CUTLER (770469).

REFERENCES

[1] Laaksonen, S., Matikainen, J., & Tikka, M. 2013. ”Tutkimusotteita verkosta” [Research approaches to internet] In Laaksonen, S., Matikainen, J., Tikka, M. (eds.) *Otteita verkosta: Verkon ja sosiaalisen median tutkimusmenetelmät*. Tampere: Vastapaino, pp. 9–33.

[2] Luduenä G. A., and Behzad M. D. and Gros C., "Exploration in free word association networks: Models and experiment” IEEE Cognitive Processing, Vol. ED-15, Issue:2, pp.195-200, 2014.

[3] Metashare.csc.fi (2019). THE SUOMI 24 2001-2014 (SAMPLE) Corpus, Downloadable Version at: HTTP://METASHARE.CSC.FI/REPOSITORY/BROWSE/THE-SUOMI-24-2001-2014-SAMPLE-CORPUS-DOWNLOADABLE-VERSION/ 14C0CBAA15B21

[4] YLE, *Sanna Ukkola: Suomi24 – vihaa kellon ympäri*. [Sanna Ukkola: Suomi24 – hate around the clock], 2015 Online. Available: http://yle.fi/uutiset/3-7856685 Accessed 14.4.2017.

[5] Ourila, J., Fiam - finnish internet audience measurement. http://fiam.fi/tulokset/. Accessed: 2018-07-30.

[6] Lagus, K.H., Ruckenstein, M.S., Juvonen, A., Rajani, C., Medicine Radar–A tool forexploring online health discussions. In: Proceedings of the Digital Humanities in theNordic Countries 3rd Conference. CEUR-WS, 2018, pp. 460–468.

[7] O. Uusitalo and M. Rökman, First foreign grocery retailer enters the finnish market, a stakeholder model, Journal of Retailing and Consumer Services, 11 (4), pp.195-206, 2004

[8] Wooffitt, R. 2005. *Conversation Analysis and Discourse Analysis: A comparative and critical introduction*. London: SAGE

[9] European Commission, 2016. *Eurobarometer 85.1. Europeans in 2016: Perceptions and expectations, the fight against terrorism and radicalization*. Brussels: TNS OPINION & SOCIAL

[10] S. Wasserman and K. Faust, Social Network Analysis. Cambridge University Press, Cambridge (1994).

[11] Newman, M. E. and Girvan, M., Finding and evaluating community structure in networks. Physical review E, 69 (2):026113, 2004

[12] J. Scott, Social Network Analysis: A Handbook. Sage, London, 2nd edition, 2000.

[13] C-X Zhai, Statistical Language Models for Information Retrieval, Morgan & Claypool 2009

[14] K. Kelton, K. R. Fleischmann, and W. A. Wallace, “Trust in digital information, ”Journal of the American Society for Information Science and Technology, vol. 59, no. 3, pp. 363–374, 2008

[15] S. Moturu and H. Liu, “Quantifying the trustworthiness of social media content,” Distributed and Parallel Databases, pp. 1–22, 2010.

[16] B. Rohrmann, “The evaluation of risk communication effectiveness,”Acta psychologica, vol. 81, no. 2, pp. 169–192, 1992.

[17] R. West, “The psychology of security, ”Communications of the ACM,vol. 51, no. 4, pp. 34–40, 2008.

[18] V. Bier, “On the state of the art: risk communication to the public,” Reliability Engineering & System Safety, vol. 71, no. 2, pp. 139–150, 2001.

APPENDIX. FULL LIST OF KEYWORDS GATHERED FROM STUDENT SURVEY

vapaus	Freedom
varmuusvarasto	Prepper stockpile
vartija	Security Guard
venäjä	Russia
verkkoturvallisuus	Computer security
vesihuolto	Water supply
viestikoelaitos	Finnish Signals intelligence organization
vihreät	The Greens of Finland
vihreät miehet	(Little) Green men
viro	Estonia
viuhkamiina	Claymore
väestönsuoja	Air raid shelter
väestönsuojelu	Protection of the populace
ydinaseet	Nuclear weapons
yksityistäminen	Privatization

Finnish	English
agitattorit	Agitators
ahvenanmaa	Åland
antti rinne	Antti Rinne
armeija	Army
aseet	Weapons
aseeton palvelus	Unarmed service
asekauppa	Arms trade
asepalvelus	Conscription
asesalakuljetus	Arms smuggling
biologiset aseet	Biological weapons
brexit	Brexit
eduskunta	Finnish parliament
EU	EU
eu-rajat	EU-borders
feminismi	Feminism
ghettoutuminen	Ghettoization
haittamaahanmuutto	Harmful immigration
harhaanjohtavan tiedon jakaminen	Distribution of misinformation
helsinki	Helsinki
humanitäärinen kriisi	Humanitarian Crisis
huoltovarmuus	Operational security (of the nation)
hybridivaikuttaminen	Hybrid influence
hyppymiina	directional anti-personnel mine
hyppypanos	Bouncing mine
hyvinvointivaltion romahdus	The fall of welfare state
hävittäjä	Fighter jet
ihmiskauppa	Human trafficing
ilmastonmuutos	Climate change
ilmatila	Airspace
ilmatila	Airspace
informaatiovääristymä	Misinformation
internet	Internet
islam	Islam
isänmaa	Fatherland
itämeri	Baltic sea
itäraja	Eastern border
jussi halla-aho	Jussi Halla-Aho
järjestelmien haavoittuvuudet	System vulnerabilities
järjestäytynyt rikollisuus	Organized crime
järjestäytynyt terrorismi	Organized terrorism
kaksoiskansalaisuus	dual citizenship
kannustinloukku	Welfare trap
kasvinsuojeluaineet	Pesticides
kemialliset aseet	Chemical Weapons
kestävyysvajo	Sustainability gap
kiina	China
kommunismi	Communism
korpisoturi	Wilderness Warrior
koulutusmalli	Education model
kriisi	Crisis
krp	National Bureau of Investigation (of Finland)
kyber	Cyber
kybersodankäynti	Cyberwarfare
kyberturvallisuus	Cybersecurity
lannoitteet	Fertilizer
luottamus	Trust
länsiraja	Western border
lääkehuolto	Medical supplies
maahanmuuttajat	Immigrants
maahanmuutto	Immigration
maamiinat	Landmines
maantie	Roads
merialue	Sea area
miina	Mine
molotov	Molotov
molotov cocktail	Molotov Cocktails
murmanski	Murmansk
nationalismi	Nationalism
nato	NATO

nettitrolli	Internet Troll
niinistö	Sauli Niinistö
näkymätön sodankäynti	Invisible warfare
omavaraisuus	Self-sustainability
pakolaiset	Refugees
pakolaiskriisi	Refugee crisis
palkka-armeija	Private Military Company
palokunta	Fire Department
panssarivaunu	Tank
patriotismi	Patriotism
peiteoperaatio	cover operation
pelastuslaitos	rescue department
persut	True Finns
pietari	St. Petersburg
poikkeusolot	emergency conditions
poliisi	police
poliisin määrärahat	police resources
politiikka	Polictics
polttoainehuolto	Fuel availability
puolustusministeri	Minister of Defence
puolustusvoimat	Finnish Armed Forces
puolustusyhteistyö	Defence co-operation
putin	Vladimir Putin
raiskaus	Rape
raja(t)	Borders
rajat	Borders
rajavalvonta	Border control
rajavartiolaitos	Department of Border control
rasismi	Racism
rauha	Peace
reservi	Military reserves
rikollisuus	Crime
ruokahuolto	Food supplies
ruotsi	Sweden
rynnäkkökivääri	Assault rifle
sairaala	Hospital
salakuljetus	Smuggling

schengen	Schengen
seksuaalirikos	Sex crime
sisu	Sisu
sisäpolitiikka	Internal politics
siviilipalvelus	Civilian service
soldiers of odin	Soldiers of Odin
sosiaaliturva	Social security
sota	War
supo	Finnish Security Intelligence Service
syrjäytyminen	Exclusion
sähkön jakelu	Electric supply
sähköverkko	Electricity network
sähkövoimala	Power plant
särmäri	
taistelukaasu	poison gas
talvisota	Winter War
tasa-arvo	Equality
teknologia	Technology
terrorismi	Terrorism
terroristi	Terrorist
tiedustelulaki	Inquiries act
tietoliikenneyhteys	Data connections
tieverkko	road network
trolli	Troll
trump	Donald Trump
tuhoontuomittu	doomed
tulli	Border control / Customs
turvallisuus	Security
turvasäiliö	Preventive detention
ulkopolitiikka	External politics
usa	United States of America
uskottava maanpuolustus	Believeable national defense
uusnatsi	Neo-nazi
vaihtoehtomedia	Alternative media
vakoilu	Spying
valeuutiset	Fake News
vapaaehtoinen asepalvelus	Optional conscription