2015/2500M-0018

ISSN 1797-3457 (verkkojulkaisu)
ISBN 978-951-25-2740-3 (PDF)

TIIVISTELMÄRAPORTTI (SUMMARY REPORT)

SemPro: Use of semantic mining and profiling in building a situational picture from Social Media and Big Data (Semanttinen louhinta ja profilointi tilannekuvan muodostamisessa sosiaalisesta mediasta sekä isosta datasta)

professor Mikko Kolehmainen, tel.044 290 2637, mikko.kolehmainen@uef.fi,
research manager Mauno Rönkkö, tel. 040 355 2202, mauno.ronkko@uef.fi,
researcher Markus Stocker, markus.stocker@uef.fi
Department of Environmental Science, University of Eastern Finland, PoBOX 1627, 70211 Kuopio

Abstract. In this project, the goal was to study, if semantic mining and profiling could be used in building a situational picture from Social Media and Big Data. Twitter was used as the primary source of Social Media. Carrot^2 service was used as a tool for extracting Big Data. Data collection and analysis methods were developed and applied to research questions that were specifically chosen and selected in cooperation with researchers working at the Finnish Defence Research Agency at Riihimäki. The research results validated that it is indeed possible to build association networks from Social Media data and Big Data. Twitter data in particular could be used to build a situational picture on a map regarding a specific discussion topic and the actors around the topic. It was further demonstrated that Social Media, in particular Twitter, supports analysis of fast events, whereas Big Data overwhelms sudden changes and discussion topics, whereby it is better suited for building long term association networks that could explain reasons for specific events. Use of co-learning was also studied in developing the association networks.

1 Introduction

The goal of using semantic mining and profiling in building a situational picture from Social Media and Big Data was challenging from many perspectives. Already the shear amount of data found in Social Media is too much for a full and comprehensive real-time, on-line analysis. For instance, in Twitter, there are already more than 200 billion messages stored. Another, obvious challenge was the language used in Social Media messaging.

The approach was to study association networks, to capture temporal and spatial correlations. Because of this, methods for temporal and spatial correlation were developed in the project. This enabled, for instance, visualization of correlations by using multidimensional graphs. The research hypothesis was that the reoccurring themes would show up in the correlations and in the graphs.

We also developed methods for analyzing and identifying relevant concepts and factors affecting the collected association networks. Such concepts and factors were then used in refining the research questions during the project. This approach made it possible to develop further the analysis algorithms used on collected Twitter data.

The co-learning aspect was also studied. The initial hypothesis in the research plan was that semi-automated methods would support co-learning, where the user could refine his or her understanding on the topic based on association networks and

visualization, and thereby also refine either the methods or the association networks to better fit the research target. We carried out the research in this manner. The methods were greatly refined during the research, thus, validating the initial research hypothesis in co-learning.

The results of this project can be directly exploited in Finnish Defence as well as in safety and security activities. The methods produce association networks and visualizations that help in understanding the complex relationships in Social Media and Big Data. In particular, we applied the methods in this project in analyzing the communication actors and their messaging in an international setting. We were able to validate that the methods do provide significant insights into the communication biases and potential factors regarding the data collected from Social Media. In particular, map based visualizations provide a situational picture that helps in understanding the spatial and temporal factors in the communication. The analysis results also validate the research hypothesis that the association networks change over time and that the changes reflect also the changes in the global relationships. Also, the analysis results clearly show how the methods are able to identify key actors that try to affect the global thinking and opinions on some specific topic.

2 Research objectives and accomplishment plan

The main research question of this project was: can semantic mining and profiling be used in building a situational picture from Social Media and Big Data? In particular, we wanted to study, if we can construct association networks that identify and explain correlations in Social Media and Big Data. During our first meeting at the Finnish Defence Research Agency at Riihimäki, we refined our topic questions to targeted questions:

- What kind of effect international crises have to Finnish Defence?
- What external actors and factors affect the discussion about Finnish Defence?
- What external actors and factors affect the discussion about Finnish-Nato cooperation?

There were several hypotheses regarding the research question:

1) Language used in Social Media and in Big Data has sufficient semantic content. The language is often colorful, biased and it contains many acronyms. Furthermore, incomplete sentences are used in the communication. Still, it is reasonable to assume that the communication has a meaningful content.
2) Reoccurring themes show up as correlations. Such themes could then be learned in a semi-automated fashion. Our approach was to study these correlations as association networks. Such networks could then be also visualized as multidimensional graphs.
3) Association network help in detecting weak signals. Detection of weak signals requires also domain expertise, whereby we do not assume that this task could be fully automated. Although, unexpected correlations can indicate potential weak signals.
4) Semi-automated methods support co-learning. In short, the user can learn on the topic based on association networks and visualization, and thereby also to guide and refine either the methods or the association networks to better fit the research target.

After the first meeting at the Finnish Defence Research Agency at Riihimäki, we refined the research plan to:

1) Implementation of the Twitter tweet collection system
2) Collection of tweets based on selected keywords
3) Collection of Big Data associations on selected keywords with Carrot2
4) Collection of BBC news feeds on selected keywords
5) Development of semantic mining methods
6) Discussions on intermediate results
7) Final analysis of Twitter data, Big Data associations, and BBC news feeds

3 Materials and methods

We used a constructive approach in the research. In short, we collected data, developed algorithms that were then applied on the data to validate our research hypothesis. Similar research method has been used before in Social Media and Big Data studies.

Regarding previous research on Social Media and Big Data, Twitter data has been used for research purposes, in particular for political orientation mining and sentiment analysis. [MR13] predict the political preference of Twitter users from their interaction with political parties. [CR13] emphasize the difficulty of classifying (political) orientation on Twitter. The authors review related published work and conclude that "reported accuracies have been systematically overoptimistic due to the way in which validation datasets have been collected.'' The authors also underscore that trained classifiers "cannot be used to classify users outside the narrow range of political orientation." [MM10] discuss the "discovery of Twitter users' topics of interest by examining the entities they mention in their Tweets." The method is developed in order to allow for "clustering and search of Twitter users based upon their interests." Using Twitter, [TBP11] study whether popular events such as the Oscars or the Olympics are associated with increases in sentiment strength (i.e. increases in expressed strength of feeling). The authors found "that popular events are normally associated with increase in negative sentiment strength and some evidence that peaks of interest in events have stronger positive sentiment than the time before the peak." [SHP10] study the factors that tend to affect the retweetability of a tweet. The approach builds on a set of 9 features, such as number of hashtags/URLs/mentions in a tweet, and PCA. Twitter data has also been used for purposes other than studying the political orientation of users and sentiment of tweets. [MB15] developed an open source toolkit for analyzing Twitter data for the purpose of understanding opinions on climate change on social media. Beyond terms, [MB15] also extract topics and sentiments. In environmental studies, monitoring of global air quality through participatory sensing has been studied over the past four years. We have also cooperated [RSR15] with professor Kostas Karatzas (Aristotle University of Thessaloniki, Greece) on using Twitter data to analyze effects of air quality on an individual level by using neural nets and clustering.

3.1 Social Media (Twitter)

We collected tweets matching 199 word pairs, such as (Ukraine, Russia) and (Greece, EU), over the period March 9 and October 5, 2015. We took a snapshot on October 5 in order to complete the analysis for this report and project. The obtained Twitter data is formatted in JavaScript Object Notation (JSON) and is persisted to disk. Persistence and management of the JSON Twitter data is taken care of by MongoDB.

The persisted Twitter data is analyzed offline. Specifically, we focus on descriptive statistics, text mining, and user profiling for specific sub-collections of the entire collection. Sub-collections match word pairs related to current crises of interest to Finland and Europe. We analyzed the collections for the (Ukraine, Russia) and (Greece, EU) pairs. We also performed the analysis for a crisis for which we did not explicitly filter word pairs, namely the Syrian crisis using the word pair (Syria, EU). With this we want to investigate if the proposed approaches are sensitive to crises other than those configured.

The evaluation was performed as follows. First, we provided some descriptive statistics of the full collection and the three sub-collections for the chosen word pairs. Descriptive statistics included the total number of tweets in the respective collection, the distinct users, and the user with maximum number of tweets, both for the collection and

the geo-located sub-collection of the collection. We also reported on the number of users with 1-20 tweets in the entire collection. Finally, we extracted the trend over time for the daily count of tweets for each collection.

Second, we performed text mining on the collections for the selected word pairs. Specifically, we computed the most frequent terms associated with each of the word pairs. Furthermore, for each collection we created topic models that show the trend of the 5 most salient topics over time, individually for each studied crisis. Text mining was performed on random samples of size 10,000.

Third, we performed a user profile analysis by clustering users into engaging and non-engaging depending on features of their tweets. Features included: original tweet length, whether tweet starts with @, number of @s, number of hashtags, number of links, number of words, normalized tweet length, number of followers, number of tweets, number of favorites.

Fourth, we developed an approach to provide a situational picture with map visualization of user profiles. We integrated various characteristics of users into the visualization. The different characteristics of user profiles were visualized using colors, size and transparency marking the location of the tweeting user on the map.

We developed two main software components: a collector and an analyzer. The collector was implemented in Java and it was responsible to collect tweets that match at least one of the 199 chosen word pairs. The collection was implemented using a persistent Twitter stream query that filters on the chosen word pairs. We used the Twitter4J Java library in communication with the Twitter API. The analyzer was implemented in R and it was responsible for automated analysis of the collections. We thus performed the aforementioned descriptive statistics, text mining, and user profile analyses in R, using RStudio.

3.2 Big Data (Carrot2)

During April and September 2015, Carrot^2 service was used to discover association networks in Big Data including Internet search results from Google and Bing, as well as content from Wikipedia and News services. In the collection process the following 21 search phrases such as: “Finland Greek Debt”, “Finland NATO”, and “Russia EU”. Searches were carried out once in two weeks during the April and September.

In each search, Carrot^2 service provided a list of associated terms and grouped them into a foam tree. The terms were then ranked into three categories depending on their centrality in the foam tree: 1) the hot terms, 2) intermediate terms, and 3) peripheral terms. For detailed analysis of how the terms and the phrases occurred in the search data, specific software was developed in Java, to map, sort, and count found terms and keywords, and to provide desired subsets of the found terms with respect to search phrases. Due to the nature of the search results by Carrot^2, the analysis method was based on temporal frequency analysis. Also, because of this, clustering of terms was done manually. This also supported the research objective of co-learning.

We also collected BBC news feeds with these search terms during the period June – September 2015. The purpose was to compare important international news headlines to association networks to determine the degree of overlap. Although Carrot^2 service includes news, it does not mean that it fully considers them. This is why (BBC) news headlines form still a complementary source for information. The comparison of news headlines to found search terms was done manually as part of the co-learning exercise.

Lastly, we also applied Carrot^2 Workbench clustering methods to tweets collected with the same phrases. This provided an alternative view to mining of common topics, terms, and themes in tweets. Also, the purpose of this task was to compare how the classification method, Lingo, that is used by the Carrot^2 Workbench and service performs with respect to the methods developed in this project for analyzing the collected tweets.

The lingo algorithm is based on use of singular value decomposition.

3.3 Literature

[CR13] Raviv Cohen and Derek Ruths. Classifying Political Orientation on Twitter: It's Not Easy! In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, 2013.
[MB15] Diana Maynard and Kalina Bontcheva. Understanding climate change tweets: an open source toolkit for social media analysis. EnviroInfo, 2015.
[MM10] Matthew Michelson and Sofus A. Macskassy. Discovering Users' Topics of Interest on Twitter: A First Look. In Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data, October 26, Toronto, Ontario, Canada, 2010.
[MR13] Aibek Makazhanov and Davood Rafiei. Predicting Political Preference of Twitter Users. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, August 29-29, Niagara, Ontario, Canada, 2013.
[RSR15] Marina Riga, Markus Stocker, Mauno Rönkkö, Kostas Karatzas, Mikko Kolehmainen. Air quality information extraction from Twitter with the use of Self-Organizing Maps. Journal of Environmental Informatics, 26(1), pp. 27-40, 2015.
[SHP10] Bongwon Suh, Lichan Hong, Peter Pirolli, end Ed H. Chi. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. IEEE International Conference on Social Computing / IEEE International Conference on Privacy, Security, Risk and Trues, 2010.
[TBP11] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in Twitter Events. Journal of the American Society for Information Science and Technology, 62(2):406-418, 2011.

4 Results and discussion

4.1 Social Media (Twitter)

The complete collection consists of roughly 8 million tweets by 1.5 million distinct users. The most active user seems to be a Russia news agency. Interestingly, the account is suspended by Twitter. In the complete collection, only 0.17% of tweets by roughly 6 thousand distinct users are geo-located. The most active user with geo-located tweets is Paul Erickson @epaulnet.

The collections for Ukraine-Russia and Greece-EU have a comparable total number of tweets, with Greece-EU having 60% more distinct users. The Syria-EU collection has significantly less tweets than the collections for the other two crises. This is arguably due to the fact that we did not explicitly filter the (Syria, EU) pair in our persistent streamed Twitter query.

The vast majority of users have merely one tweet matching any of the configured word pairs. The number of users with at least five tweets drops dramatically so that the number of users with at least 20 tweets matching a word pair is comparatively extremely small.

For the complete collection and collection period, there is a considerable increase in activity during July with three peaks lasting a couple of days. Regarding the three selected crises, Ukraine-Russia, Greece-EU, and Syria-EU, there are three distinct peaks in the number of tweets. For the Ukraine-Russia crisis, the tweet volume has a slight downward trend over the period suggesting that the Ukraine-Russia crisis had lost some of the saliency over the summer months 2015. For the Greece-EU crisis, there are three July peaks that can arguably be attributed to the news about the desolate state of Greece in July, in particular also the government decision to restrict access to banks. For the Syria-EU crisis, the volume of tweets has increased in early September, which coincides

with the events and news about Syria and the refugee crisis in Europe.

For the three crisis of interest to Finland and Europe, we performed simple text mining tasks, in particular (1) analysis of the most frequent words related to, and (2) topic modelling for, each crisis.

As for the most frequent words in the Ukraine-Russia collection, they are: world, war, usa, terrorrussiaukrain, stoprussianaggress, putinkil, putin, nato, military, crimea, banrussiafromswift. The worlds clearly associate with the Ukraine-Russia crisis. As for the topic model for the Ukraine-Russia crisis, five topics were identified: crimea, nato, usa, war, and world. The topics were mined as relevant by an algorithm centered on Latent Dirichlet Allocation (LDA), a generative model in natural language processing.

As for the most frequent words in the Greece-EU crisis, they are: vote, leave, imf, germani, eurozon, European, europ, ecb, debt, deal, bank, bailout. The five identified topics were: deal, germani, russia, talk, vote. Interestingly, the topic model clearly shows the July peak, suggesting that the crisis is media driven.

As for the Syria-EU crisis, the most frequent words were: ukrain, russia, putin, pope, overburden, obama, isi, insid, exodus, drown, differ, child, beltway, assad. The discovered topics were: assad, child, axodus, isi, ukrain. Interestingly, the topic “assad” seems to be the dominant topic in the recent increased saliency of the Syria-EU crisis since late August. Also notable is that the approach was able to detect the world-wide viral news of the three-year-old who was found dead washed up onshore on a beach.

The third analysis of Twitter data was an attempt to profile users and provide an informative situational picture of users, their classification, activity, and relevance on Twitter. For the purpose here, we classified user in either engaging or non-engaging based on structural features of their tweets. Together with tweets and retweet volumes, and tweets geo-location, this classification was used in a tool that visualizes user profiles on a map. This visualization provides a situational picture, designed for quick and easy digestion of large amounts of Twitter data by operators. Given a map it is, for instance, straightforward to identify non-engaging users who tweet a lot but are seldom retweeted. The expert-tool interaction is an example of co-learning.

In the user clustering, the hypothesis was that engaging users more often refer to other users by mentioning them in tweets and use the 140 characters to write as content-rich a tweet as possible. Hence, for the engaging users we expect more words and longer (normalized) tweet texts as well as less hashtags and links. The clusters were well separated. Given this interpretation of clusters, we created map visualizations for the situational picture of user profiles. Specifically, we mark the location of users differently based on if the user was engaging or not, and on the number of tweets and retweets. The visualization was then applied to geo-located tweets. For the Ukraine-Russia crisis, it appears that most users are not engaging, tweet relatively little and only few are retweeted. For the Greece-EU crisis, we noted a relatively strong density of engaging users in the UK and one user who tweets a lot, is not often retweeted, and is not engaging in Greece.

4.2 Big Data (Carrot2)

The total number of uncovered unique terms associated with the search phrases during the period of April - September 2015 was 3182. The average number of terms per phrase was 145 and the standard deviation was 16. This was unexpected. During the collection, the number of new unique search terms drops quickly from about 25 to about 10 for each phrase after some searches. There is a clear increase in the number of new unique terms in the next collection in August due to holiday break. These observations indicate that new topics and events are undermined simply by the mass of existing data regarding past events, ruling out use of Carrot[2] service as part of real-time analysis.

We performed simple trend analysis of found term per each search phrase. We

found some interesting similarities. For instance, the trends for the phrases “Ukraine Putin” and “Finland Greek Debt” were unexpectedly very similar. In this sense, the trend analysis can be used as an indicator. It shows correlations between specific search terms, and thus, apotential existence of common factors and causalities.

For all search phrases, we collected the found terms and their ranking. For instance, for a search phrase “Russia EU”, we found the terms “Oil” and “Sanctions against Russia” with ranking 1. With ranking 2, we found terms, such as: “Crisis in Ukraine”, “Energy Cooperation”, “Trade war with Russia”, and “Vladimir Putin”. With ranking 3, we found terms, such as: “Annexation of Crimea”, “Brussels”, “Dairy Products”, “Interested in Modern Russia and the Future”, “Russia's Gas Pipeline Strategy and Europe's Alternatives”, and “Russia's Travel Ban”.

We formed association networks for selected search phrases and terms. For instance, for “Russia EU”, the association network was formed before June 2015 and late September 2015. We observed a clear growth of the association network. The latter network included most recent events and made the association connections clearer. The resulted association networks, however, indicated that the collection period of eight months is too short for analytic purposes. Still, the association networks do indicate that they develop over time by gaining more details and accuracy.

We also analyzed, how the ranking of terms evolved during the collection period. We found out that the ranking changes can be used to uncover the importance and timeliness of the term with respect to the search phrase.

We collected BBC News headlines and compared them with the collected terms for the search phrases. What was very clear, when collecting BBC News headlines, was that Finland and events related to Finland are almost non-existent. The news focus only on events having an impact on UK, the whole Europe, or a significant part of the Europe. In more detailed analysis, we found out that “Oil”, “Gas”, and “Energy Cooperation” are not as current topics for (BBC) news as some other topics. This could reflect the emphasis of BBC News on matters that have significance on UK. Some topics, such as “Ukraine Crisis” were consistently covered a longer period of time. We also found that there were terms in BBC News that were of (news) value, but did not appear among the search phrase, yet. Syria was one such topic during the collection.

To provide a comparative view to the methods developed in this project on analysis of Twitter data, we applied Carrot^2 Workbench on collected tweets. We applied the workbench and its clustering methods to subsets of tweets collected with a specific phrase. The tweets were merged based on the sender. In general, all the results show similar kind of terms as were obtained from Big Data. Interestingly, the found terms mirrored quite well the terms found through the Carrot^2 service. However, more detailed terms were found as more tweets were used. Still, with 4000 tweets, repetition of terms with slightly different wording became apparent. This means that Carrot^2 Workbench should be used on relatively small subsets of tweets.

4.3 Exploitation of the results

We have analyzed the following three international crises: Ukraine-Russia, Greece-EU, and Syria-EU. The developed methods on Twitter data are useful to obtain a situational picture for important topics relevant to the crises, and their development over time. Analyzing Twitter data can provide clues for what could become threats based on what people discuss on Twitter. Finnish Defense can use the developed methods in urgency assessment.

The key topics emerging from Twitter data for the Ukraine-Russia crisis underscore the weight of certain actors in the crisis, such as NATO and the USA. These actors traditionally dominate the discussions, negotiations, and, ultimately the course of history. Specific keywords found in twitter data, such as “banrussiafromswift”, also underscore

the effects of non-military action, in this case economic effects. Economic sanctions affect international trade and disproportionally also Finland.

Compared to the Ukraine-Russia crisis, Greece-EU has obviously a very different character, more economic than military. This can clearly be seen in the emerging keywords (e.g. “imf”, “ecb”, “bank”, and “bailout”). The effects of the Greece-EU international crisis are thus of more interest to the Finnish government at large than to Finnish Defense, specifically.

The Syria-EU is clearly again a military crisis. Compared to the Ukraine-Russia crisis, there is however an important difference, namely the refugee crisis. It is obviously largely a consequence of the ongoing war in Syria and has implications to Finnish national security. That the refugee crisis is an important dimension of the Syria-EU crisis is reflected in the emergence of the keyword “exodus”, as well as “drown” and “child” which reflect some of the tragic events.

Twitter user profiling and the situational picture proposed in this work is an interesting approach to individuate potentially interesting external actors affecting discussion on Finnish Defence. Specifically, the proposed approach can individuate individuals or organizations that are particularly vocal on a specific topic, fall into a certain category, and are located in areas of interest. For the studied crisis, we found that it is often news agencies and sometimes individuals (in particular journalists) that emerge as the most vocal actors. Given the automatically generated geospatial situational picture, it is however up to an expert to further investigate the relevance of an actor to a specific goal. The resulting interaction between computational and human agents facilitates co-learning.

We shall next discuss how the results on Social Media and Big Data relate to the topical questions posed at the beginning of the project.

<u>*What kind of effect international crises have to Finnish Defence?*</u>

1. *The borderline of Finland becomes a focal topic.* The collected terms include discussion around Finnish borderline. Moreover, such a discussion seems to correlated with the discussions revolving around neutrality of Finland with respect to NATO and relation of Russia to Baltic states.
2. *Effects on weapons acquisition.* The collected terms include weapons whenever there is a crisis. This term could be seen to be coupled with terms related to a political debate and alliances as well as sanctions. Together all this can be seen to indicate potential effects on weapons acquisitions in case of crisis.
3. *Relations to alliances and organizations.* In case of crisis, Big Data indicates that there are always alliances, agreements, and global organizations, such as NATO, affecting the strategies during crisis. This could be seen to indicate that that modern defence is not in a vacuum, isolated from an international political scene.
4. *Participation in support actions.* The relations to alliances and organizations require and/or bring in support actions. Trade wars and sanctions are such examples.
5. *Effect of trade wars and sanctions.* Based on the found terms, there is an indication of an indirect effect of crisis to Finnish Defence through commerce. As confirmed by news, trade wars and sanctions affect directly the availability of specific products and services, which has a negative effect on society and economy over a longer time period.
6. *Energy availability and self-sufficiency.* The terms related to energy and oil seem to correlate with crisis. This could be seen as an indication of the significance of energy availability and self-sufficiency during crises.
7. *Public relations and international debate.* The BBC News headlines strongly indicate that Finland is not really appearing in international scenes. At the same time, as suggested by the found terms, propaganda and “information warfare” is a current means to influence the political scene during a crisis.

What external actors and factors affect the discussion about Finnish Defence?

1. *International affairs.* The found terms indicate that the relationship between EU, Russia, and NATO dominate the discussion about Finnish Defence.
2. *EU actions.* The found terms also indicate that EU actions are also a significant factor to discussion on Finnish Defence. In particular, sanctions, travel bans, and trade wars, reflect also in the discussions.
3. *Lack of shared international interests.* The BBC News headlines strongly indicate that the lack of shared interest by some country (such as UK) also show as lack of interest to shared discussions.
4. *Individuals.* The found terms indicate that the discussion is somehow rooted to internationally strong political individuals.
5. *Propaganda.* This was a found term in the discussions.
6. *Elections.* This was another found term in the discussions.
7. *Geographics.* The Finnish borderline and arctic dimensions were among the found terms indicating of the unique aspects of Finland.
8. *Energy security.* Among the found terms, there were terms such as oil, gas, and energy, indicating that energy security is also linked to defence.
9. *Refugee crisis.* The term also appeared among the collected terms. Its significance on European level has recently been highlighted on daily news.

What external actors and factors affect the discussion about Finnish-NATO cooperation?

1. *Russia.* Found terms indicate that Russia's relationships in Europe and especially to Finland are a dominant factor.
2. *Ukraine crisis.* The found terms include actions in Ukraine. There are also correlations between Ukraine and NATO as found terms.
3. *EU relationships.* EU and its actions regarding Europe and other countries did also show up in found terms.
4. *Baltic States.* Baltic States appeared either as individual countries or as a uniform region among found terms.
5. *Image of neutral Finland.* Another found term in the discussion about Finnish-NATO cooperation was "neutral Finland".
6. *Sweden.* Sweden's relationship to NATO was also found among the terms as a factor in Finnish-NATO discussions.
7. *Politicians and elections.* "Elections" was found both as a term on its own among the collected terms. In particular, Finnish president and prime minister were found among the terms as a significant factor.
8. *Military exercises.* Among the collected terms, "Military exercises" as well as other similar terms were found as factors.

5 Conclusions

The aim of this project was to investigate whether semantic mining and profiling can be used to build a situational picture from Social Medial and Big Data. The developed approaches were used to address questions regarding the effects international crises have to Finnish Defence and the external actors and factors that affect discussions about Finnish Defense and the Finnish-NATO collaboration.

Social Media focused on Twitter and data collected from Twitter. We computed summary statistics, performed text mining, and user profiling to obtain a situational picture. We also developed an approach for situational picture visualization of users and their geo-location, classification into engaging and non-engaging, and tweet and retweet volumes. The resulting visualization can be used as a tool in co-learning processes,

whereby software agents create from data a situational picture for experts who can then further analyze the information presented visually.

We observed that text mining can be used to draw relevant terms and phrases associated to one or more phrases of interest. Topic modelling can discover focal topics and their trend over time. User profiling can condense the large amount of data into easy to digest information, intuitively mapped for geo-location and user Tweet characteristics.

As for the Big Data, we found that the Carrot^2 service and the methods for analysis implemented in the Carrot^2 service support building of association networks based on the information available through Google and Bing, as well as content from Wikipedia and News services. The networks are then, however, build over a long period of time, as the often discussed topics overwhelm the new topics in search results.

An active monitoring of news headlines is still required, to understand the context for found associations. Then, however, the source for the news has a subjective bias that should be taken into account. This was very apparent, when BBC News headlines were used as a background material to understand events related to Finland. In particular, this very case highlighted how little relevance Finland has for UK based on very limited number of news headlines discussing Finland or events related to Finland.

Although the development of association networks through Big Data requires fairly long period of time, and the results are not well suited for identifying weak signals, the results do provide information that is useful for assessing, for instance, the effect of international crises and external factors to Finnish Defence.

There exist several directions for future work. First, the large data volumes require state of the art Big Data analytics data management and processing systems. Of interest are systems such as Apache Hadoop or Apache Spark. Big Data analytics tries to avoid sampling and instead analyze the complete dataset, which is particularly important to discover weak signals. Second, it may be interesting to provide the situational picture, and other analytics, on streamed data. Future work could thus investigate the application of the approaches discussed here to streamed data. Third, the discussed approaches and methods can be developed further. For instance, feedback from users could be valuable for the development of the tool for situational picture. Fourth, better collaboration with Finnish Defense forces could lead to further interesting results. Fifth, semantic technologies could be introduced in this work to formally describe users and their relations with spatio-temporal and thematic dimensions. The resulting knowledge base could support flexible and dynamic querying and thus retrieval and discovery of information by agents, in particular experts. Such a knowledge base could populate a visual interface but would also allow for other types of programmatic interaction with actionable knowledge extracted from Twitter data.

6 Scientific publishing and other reports produced by the research project

During this project, we have not published the results, yet. A manuscript for a scientific article is planned to be written after project ends.

The collected Twitter dataset is to be used in further studies. However, the Twitter license prohibits the distribution of the data as such. Therefore, the data is available only through research cooperation.