Analyses of information security standards on data crawled from company web sites using SweClarin resources

Arne Jönsson
Computer and Information Science
Linköping University
Linköping, Sweden
arne.jonsson@liu.se

Subhomoy Bandyopadhyay
Management and Engineering
Linköping University
Linköping, Sweden
subhomoy.bandyopadhyay@liu.se

Svjetlana Pantic Dragisic
Management and Engineering
Linköping University
Linköping, Sweden
svjetlana.pantic.dragisic@liu.se

Andrea Fried
Management and Engineering
Linköping University
Linköping, Sweden
andrea.fried@liu.se

Abstract

With the purpose of analysing Swedish companies' adherence and adoption of the information security standard ISO 27001 and to examine the communicative constitution of preventive innovation in organisations, we have created a dataset of corporate texts from Swedish company websites. The dataset is analysed from multiple interdisciplinary perspectives in close cooperation with organisation researchers and SweClarin researchers using SweClarin tools and resources as well as standard language technology tools. Some analyses require deep reading, which is performed by organisational studies researchers. Initial results have been presented at an organisational studies conference. In this paper, we focus on presenting the research issues, the methods used in the project, and the experience of SweClarin researchers supporting researchers in social sciences. Our contribution is to show how it is possible, through triangulation of human and digital methods, to increase the credibility and validity of a digitally acquired data set and subsequent research findings. In our view, a combination of human deep reading (organisation researchers), contextual dictionary verification (organisation and management studies) and language technology (sentiment analysis) can help to sensitise computational text analysis for medium-sized data sets.

1 Introduction

Preventive innovation differs from ordinary innovation. The innovation literature claims that the economic benefits of preventive innovation to organisations, for instance, for avoiding environmental pollution, protecting human health or ensuring information security, are mainly intangible, often time-delayed and adopted for incidents that may never occur (Rogers, 1995). To address these challenges, organisational communication seems to be crucial to increase the potential of economic recognition for preventive innovation.

Therefore, we draw on the "communicative constitution of organisations" view to explore how preventive innovations are communicatively constituted. Using the example of the information security standard ISO/IEC 27001, we examine how communication of preventive innovations is shaped by its adopting organisations. We analyse texts about the information security standard ISO/IEC 27001 on Swedish corporate websites supported by computational tools for web scraping and language analyses. As a result, we first identify three communicative practices of data governance termed agency, stewardship and brokerage, and second, provide evidence that organisations' communication also depends on whether they receive direct or indirect economic recognition for their preventive innovation.

We contribute a meaningful combination of deep reading of humans (researchers), dictionary verification for a specific context (innovation research) and language technology (sentiment analysis) to a meaning-centred and situational understanding of preventive innovation. Our analysis enhances Rogers' perspective by challenging the classification of preventive innovations as mere ”isolated, static objects or practices”, unveiling their dynamic interplay with organisational members — simultaneously influencing and being influenced — i.e., are enacted communicatively by organisations. Contrary to Rogers' assumption, we also provide initial evidence that preventive innovations can very well achieve economic recognition by constituting different meanings of preventive innovation.

This paper will focus on the methodology, rather than delving into the theoretical underpinnings. We illustrate the potential of SweClarin and language technology analyses for investigating organisational communication and the production of meaning in their texts.

2 Research design

Using ISO/IEC 27001 as an example to study the communication of preventive innovations, our research design followed three steps, see Figure 1. We first generated a dataset of Swedish corporate websites of all sectors and scraped the content for ISO/IEC 27001 related paragraphs of the text corpus. Second, we categorised the identified companies manually according to their adoption (of preventive innovation) approach. Finally, we conducted analyses on the language used in the paragraphs relating to ISO/IEC 27001 on these websites.

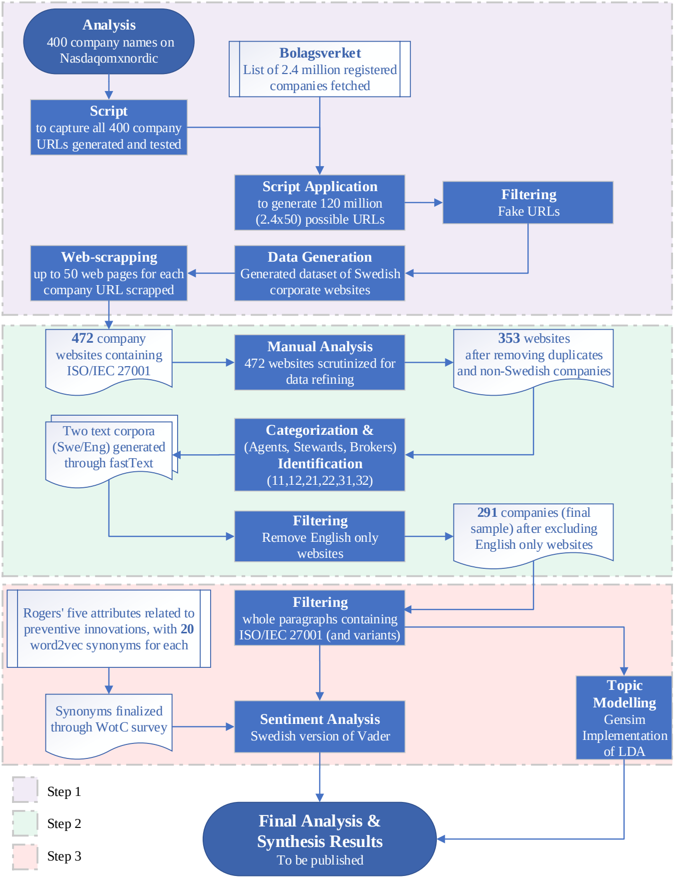
Analysis
400 company names on Nasdaqomxnordic
Bolagsverket
List of 2.4 million registered companies fetched
Script
to capture all 400 company URLs generated and tested
Script Application
to generate 120 million (2.4x50) possible URLs
Filtering
Fake URLs
Web-scrapping
up to 50 web pages for each company URL scrapped
Data Generation
Generated dataset of Swedish corporate websites
472 company websites containing ISO/IEC 27001
Manual Analysis
472 websites scrutinized for data refining
353 websites after removing duplicates and non-Swedish companies
Two text corpora (Swe/Eng) generated through fastText
Categorization &
(Agents, Stewards, Brokers)
Identification
(11,12,21,22,31,32)
Filtering
Remove English only websites
291 companies (final sample) after excluding English only websites
Rogers' five attributes related to preventive innovations, with 20 word2vec synonyms for each
Filtering
whole paragraphs containing ISO/IEC 27001 (and variants)
Topic Modelling
Gensim Implementation of LDA
Synonyms finalized through WotC survey
Sentiment Analysis
Swedish version of Vader
Step 1
Step 2
Step 3
Final Analysis &
Synthesis Results
To be published

Figure 1: Overview of the process.

Regarding the first step, as a complete dataset of all websites of Swedish companies does not exist as open access, we contacted several institutions to retrieve this data. We approached Sweden's company registration office, Bolagsverket, and Statistics Sweden (SCB) to get access to company names, identification numbers, sector affiliations and innovation indicators. However, Bolagsverket and SCB could not provide a database with company URLs. We, therefore, analysed 400 company names on Nasdaq Nordic (https://www.nasdaqomxnordic.com/) through scripts that generate web addresses in order to understand how company URLs can be constructed, and used that to generate 120 million possible URLs from the 2.4 million registered companies listed on Bolagsverket. These URLs were then tested to check how many of them were actual websites. These websites were then scraped in September 2020. We scraped up to 50 connected web pages of each site to grasp sufficient content (cf., Kinne and Lenz (2019)). Out of all scraped websites, we found 472 which contained the filter phrases 'ISO 27001', 'IEC 27001', 'IEC 270' or 'ISO 270'[1].

After we had identified the 472 websites[2], as a second step, we manually analysed each company's website by visiting their URLs to verify the scrapped data. This hands-on scrutiny of the corporate websites aimed to refine the extracted information regarding companies' certifications, business sectors, models, and value propositions. After removing duplicates and further non-Swedish companies in the dataset, we were left with 353 websites of Swedish companies. We categorised these companies according to the criteria 'certified' or 'non-certified', following a suggestion by Mirtsch et al. (2020). Their findings reveal that a third of the companies that adopt ISO/IEC 27001 do so through certification, with the remainder opting for non-certified pathways. Furthermore, our findings revealed a variety of companies: some integrated ISO/IEC 27001 consulting or training into their business models, while others, lacking certification and refraining from offering consulting or training services, solely referenced certified clients, customers, and suppliers on their websites. Based on this initial categorisation, we identified six distinct types of preventive innovation adoption, denoted as 11, 12, 21, 22, 31, and 32.[3] into which each company belongs.

In addition, two text corpora were generated from all identified company websites, one in Swedish and one in English. We use fastText (Joulin et al., 2016) to separate the sentences. For each company, we take each sentence and place it in an English or a Swedish text file, i.e. a company can have two files, one with English text and one with Swedish. The English text corpus, spanning around 450 pages, underwent manual analysis through deep reading, revealing that over 50% of the dataset consisted of inconsequential noise such as ads, menus, contact details, and website cookies. As an outcome of this analysis and in pursuit of methodological rigour through Swedish sense-based sentiment analysis (elaborated upon below), companies with English only websites were excluded from the sample, resulting in 291 companies (final sample size)[4] with websites in either Swedish or both Swedish and English. Although certain Swedish websites maintained English versions, it is noteworthy that, for analytical efficiency, the term ”Swedish only” pertains solely to these 291 entities, since the English text corpus had been excluded from further analysis. Table 1 depicts the number of sentences and words for each adoption type for the 291 companies with Swedish only text on their web pages.

This resulted in a text corpus of close to 9 million words, see Table 1. Content analysis on these texts was performed to demonstrate how preventive innovation is manifested within the communication of the six identified adoption approaches. To aid this analysis word clouds were created using the WordCloud package[5] and topic analysis using the Gensim implementation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003). We also translate all Swedish texts to English using googletrans[6], as not all organisational studies researchers are fluent in Swedish.

To assess the relevance and usefulness of preventive innovations along five attributes (as suggested

[1] Including variants such as iso-27001 and Iso 270.

[2] Available at https://www.ida.liu.se/~arnjo82/472_webpages

[3] The first digit (1, 2, or 3) denotes the three data governance approaches: Agents, Stewards, and Brokers, whereas the second digit (1 or 2) signifies (in)direct economic benefits resulting from preventive communication adoption, evaluated based on ISO/IEC 27001 training/consultation provision.

[4] Available at https://www.ida.liu.se/~arnjo82/291_filtered_webpages

[5] https://pypi.org/project/wordcloud

[6] https://pypi.org/project/googletrans

Adoption type	11	12	21	22	31	32
Number of companies	103	10	81	41	19	37
Number of sentences	197.131	11.225	127.880	29.404	20.462	82.390
Number of words	3.374.348	187.630	2.133.516	547.543	351.837	1.683.044
ISO paragraphs	520	88	401	133	17	372
Sentences in ISO paragraphs	8356	1153	4404	2248	561	38817

Table 1: Descriptive statistics for the Swedish companies in each adoption type

by Rogers), we use sentiment analysis. We want to compare the overall sentiment for each attribute and also compare the sentiment when ISO/IEC 27001 is presented. The paragraphs in the files containing 'ISO/IEC 27001', and its possible variants, were filtered out of each text to be used for sentiment analysis. We use the context in which an ISO/IEC 27001 sentence occurs, i.e. the whole paragraph, as it is scraped from the web. This filtering resulted in a considerably smaller number of paragraphs and the sentences within them (Table 1).

We use sentiment analysis along five attributes: relative advantage, compatibility, complexity, trialability and observability (Rogers, 1995). To capture various uses of the attributes, synonyms were generated for each attribute by using the Gensim package Řehůřek and Sojka (2010). For each attribute we generated 20 synonyms using seeds, in Swedish, that reflected the various attributes. For three of the attributes, we generated a second set of synonyms using different seeds. The general applicability of these twenty computer-generated synonyms in the Swedish colloquial language was assessed through a wisdom-of-the-crowd (WotC) survey approach (Surowiecki, 2004). An online Microsoft Forms survey with these twenty synonyms was sent to native Swedish speaking innovation and entrepreneurship researchers at Linköping University to compile a final set of synonyms for the five attributes.

For sentiment analysis, we use a Swedish version of Vader (Hutton & Gilbert, 2014) that considers a word's sense. Vader is a lexicon and rule-based sentiment analyser. The lexicon in English Vader comprises 5500 lexical entries with sentiment scores between +5 and -5. We used the Swedish SenSALDO 0.2 sentiment lexicon (Rouces et al., 2019) with sentiment scores -1, 0 and +1. SenSALDO 0.2 comprises 12287 lexical entries of which 8893 are unique words. Word sense disambiguation with the SenSALDO 0.2 lexicon is achieved by first parsing the texts using the Sparv pipeline[7] (Borin et al., 2016). Vader also uses booster words, such as amazingly, to further refine the sentiment analysis. The booster dictionary used in these analyses is a enhanced version of the Swedish dictionary used for sentiment analysis of e-mail conversations (Borg & Boldt, 2020) and comprises 89 items. The version used in this project, using the SweClarin SenSaldo resources, has also been used in a project on analysing Swedish official texts (Ahrenberg et al., 2022).

Data from websites are very noisy containing repetitions, menu items, contact information, adverts, etc that need to be handled. Standard crawling packages provide some cleaning of the texts but there is still much that is, for instance, not syntactically correct. Despite this, we find that the SweClarin Sparv pipeline is robust and provides an analysis that can be used by the sentiment analyser.

3 Conclusions

In this study, we started with the idea of reviving the concept of preventive innovation given the attention this type of innovation is receiving nowadays. We have chosen to explore the adherence and adoption of the ISO/IEC 27001 information security standard as an example of preventive innovation addressing cybersecurity risks as one of the great challenges of our time. Using web scraping tools and computational linguistics (and content analysis on top of that), we were able to extract and analyse large amounts of text. These texts on preventive innovation ISO/IEC 27001 include communicative efforts published on

[7] https://spraakbanken.gu.se/sparv/#/sparv-pipeline

the websites of companies operating in Sweden, telling us about the way these companies are adopting the standard. We have identified different adoption approaches and related modes of data governance. These results also help us understand that the original concept as introduced by Rogers (1995) needs to be improved in terms of opportunities to derive economic benefits from preventive innovation. By relating the adoption approaches to the different modes of data it could be shown that a meaningful adoption of preventive innovations can already take place at an early stage.

The close cooperation between the organisational studies researchers and the SweClarin language technology researchers has been imperative for the success of this project. Based on the needs of the organisational studies researchers' various analyses have been performed, and assessed. It was, for instance, initially assumed to be important to use topic models to guide the deep readings. The topic maps, however, turned out to be rather diverse and did not form a clear picture of the various adoption types. Instead, word clouds were developed that gave a better, but not sufficient, analysis of the data. In further discussions with the organisational studies researchers, we decided to try sentiment analysis, which turned out to give useful results on its own as well as the possibility to generate quotes from the texts for each sentiment ranked by its score. This gave organisation researchers the opportunity to see a quantification of the meanings to further aid the deep readings.

References

Ahrenberg, L., Holmer, D., Holmlid, S., & Jönsson, A. (2022). Analysing changes in official use of the design concept using sweclarin resources. *Proceedings of the 2022 CLARIN Annual Conference*.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Borg, A., & Boldt, M. (2020). *Using vader sentiment and svm for predicting customer response sentiment, expert systems with applications* (Vol. 162). Elsevier.

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., & Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17-18 November, 2016*.

Hutton, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14)*.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Kinne, J., & Lenz, D. (2019). *Predicting innovative firms using web mining and deep learning* (tech. rep.). ZEW Centre for European Economic Research, Discussion Paper. 01/2019 (19-001). http://ftp.zew.de/pub/zew-docs/dp/dp19001.pdf

Mirtsch, M., Kinne, J., & Blind, K. (2020). Exploring the adoption of the international information security management system standard iso/iec 27001: A web mining-based analysis. *IEEE Transactions on Engineering Management*, *68*(1), 87–100.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora [http://is.muni.cz/publication/884893/en]. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Rogers, E. M. (1995). *Diffusion of innovations*. The Free Press.

Rouces, J., Tahmasebi, N., Borin, L., & Eide, S. R. (2019). Sensaldo: Creating a sentiment lexicon for Swedish. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 4192–4198.

Surowiecki, J. (2004). *The wisdom of crowds*. NY, NY: Anchor.