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 corpus of corporate texts from Swedish company websites. The corpus was analysed from multiple interdisciplinary perspectives in close cooperation with management researchers and SweClarin researchers using SweClarin tools and resources as well as standard language technology tools. Some analyses require deep reading, which was performed by management researchers, often guided by results from language analyses. Initial results have been presented at a management studies conference. In this paper, we focus on presenting the research issues, the methods used in the project, the results, and the experience of SweClarin researchers supporting researchers in social sciences. Our contribution is to show how it is possible, through the integration of human insights 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 (management researchers), contextual lexical verification (management studies) and language technology (content and sentiment analysis) can help to sensitise computational text analysis for medium-sized data sets.

1 Introduction

Today's organisations are increasingly compelled to adopt preventive innovation to tackle pressing issues. 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). Conversely, innovations that are not preventive tend to yield immediate benefits and results in the near term. For example, implementing the ISO 9001 standard for quality management can significantly boost operational efficiency, leading to better customer satisfaction and reduced cycle times and inventory levels (Lo & Chang, 2007). Organisations face a significant challenge balancing the need for preventive innovations, which address long-term concerns, with the immediate demand for tangible results (Fineberg, 2013). This tension between long-term strategic security and short-term financial accountability complicates decision-making. As a result, companies often find themselves at a crossroad, struggling to allocate resources in a way that satisfies both current demands and future necessities. To address these challenges, organisational communication is crucial to increase the potential of economic recognition for preventive innovation.

Therefore, drawing theoretically on the discourse perspective of organisational communication (Orlikowski & Yates, 1994; Yates & Orlikowski, 1992) we develop a communicative approach that enables a situation- and meaning-centred understanding of preventive innovation. Using the example of the information security standard ISO/IEC 27001, which is designed to mitigate future risks such as phishing emails, exploitation of stolen credentials, and software vulnerabilities, 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 (content and sentiment analysis) to a meaning-centred and situational understanding of preventive innovation. A similar approach was used by Saura et al. (2023) where they used textual analysis, sentiment analysis, and topic modelling to analyse social media from an open innovation perspective. For a systematic overview of social media analyses of innovation management see Geissinger et al. (2023).

Our analysis enhances Rogers' (Rogers, 1995) 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 findings. We illustrate the potential of SweClarin and language technology analyses for investigating organisational communication and the production of meaning in their texts.

2 Conceptual Background

The diffusion of innovation theory fundamentally revolves around communication, outlining how innovations spread within a social system's participants (Rogers, 1995). Studies highlight the critical role of creating meaning in communication about innovations to facilitate their spread and acknowledge innovative contributions. Innovation literature highlights specific characteristics that both shape how organisations communicate about adopting innovations and assist in analysing an organisation's perception of these innovations, known as secondary innovation attributes (Downs et al., 1976; Tornatzky & Klein, 1982). In examining how organisations communicate about ISO/IEC 27001, we utilise Rogers (1995) five distinct attributes that adopters use to perceive and ascribe meaning to preventive (and other) innovations: relative advantage, compatibility, complexity, trialability, and observability.

Relative advantage is "the degree to which an innovation is perceived as being better than the idea it supersedes" (Rogers, 1995, p.212). Compatibility is "the degree to which an innovation is perceived as consistent with the existing values, past experiences, and needs of potential adopters" (Rogers, 1995, p.224). Complexity refers to "the degree to which an innovation is perceived as relatively difficult to understand and use" (Rogers, 1995, p.242)). Trialability is about the extent to which the innovation can be experimented with by adopters (Rogers, 1995), while observability covers the ability to see, imagine, or explain the innovation's results to others (Rogers, 1995).

To advance a situational and meaning-centred perspective on preventive innovation, we employ these attributes and theoretically anchor our research in the discourse perspective of organisational communication (Fried et al., 2024). This perspective emphasises communication's role in community development, such as the Swedish companies focusing on preventive innovation for information security, cf. Putnam (1999). Discourses, as communicative acts, form and reflect innovation's essence and outcome, shaping and reshaping innovation's meaning within organisations, cf. Bergquist (1993), Putnam (1999), and Taylor (1993).

3 Generating the corpus

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.

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 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 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 paragraphs. For each company, we take each paragraph 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. Examples of paragraphs from

[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

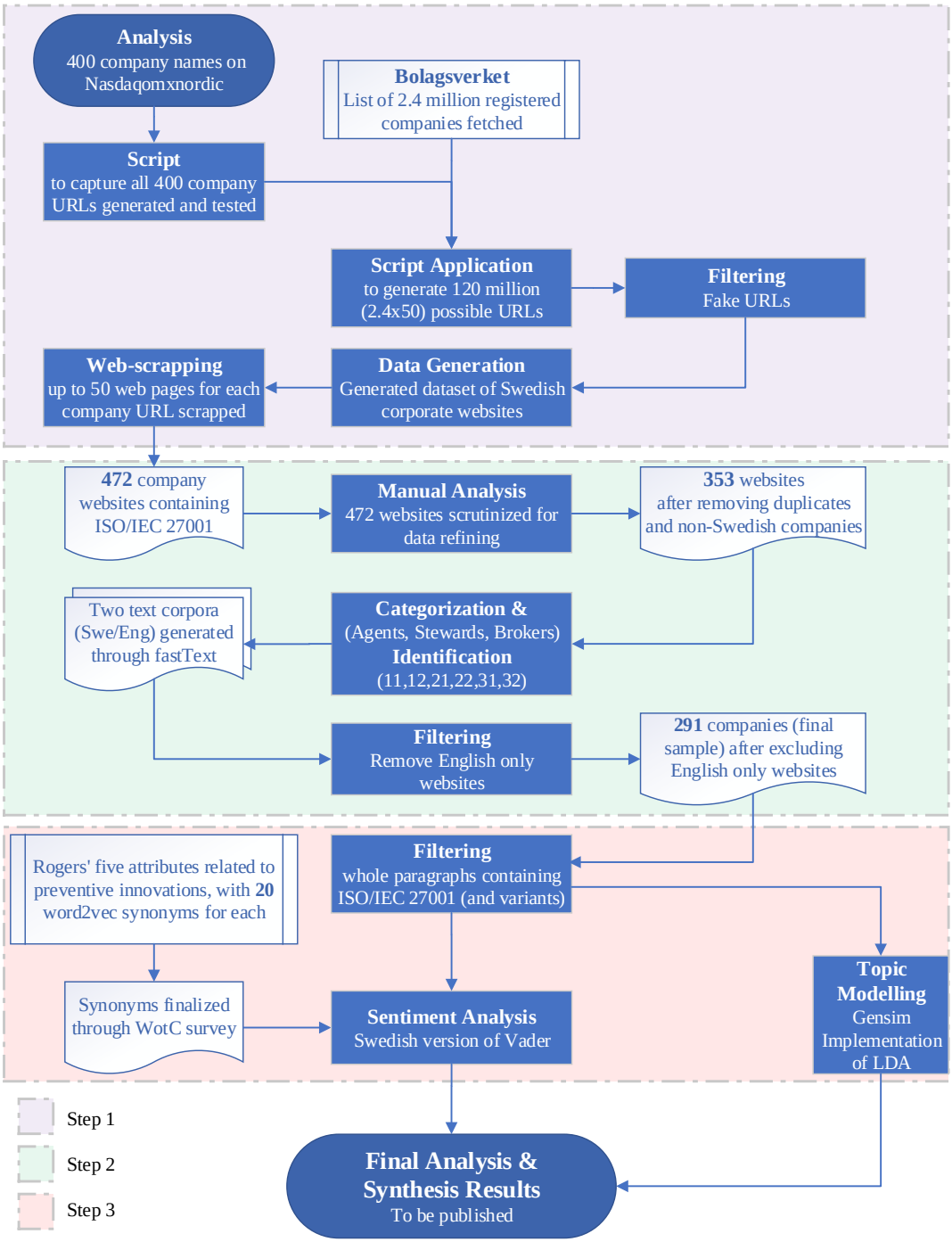

Figure 1: Overview of the process.

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	8,356	1,153	4,404	2,248	561	38,817

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

the corpus can be seen in Figure 6. We translate all Swedish texts to English using googletrans[5], as a few management studies researchers are not fluent in Swedish. All texts are also parsed using the Sparv pipeline[6] (Borin et al., 2016) to give us lemmas, POS and word sense information, needed in our analyses.

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. We identified more than 400 sentences with more than 250 words, which illustrates the noisiness of the data, in this case, the lack of proper sentence delimiters. We therefore filtered out sentences with more than 300 words and paragraphs with more than 500 sentences which resulted in a slightly smaller corpus, see Table 2.

Adoption type	11	12	21	22	31	32
Number of sentences	186,881	10,679	120,194	27,6624	19,246	77,482
Number of words	3,181,830	182,416	1,957,894	501,247	320,316	1,608,769
ISO paragraphs	373	77	203	79	10	168
Sentences in ISO paragraphs	5,943	1,088	3,028	1,620	349	16,533

Table 2: Descriptive statistics for the Swedish companies in each adoption type with sentences containing more than 300 words and paragraphs with more than 500 sentences removed

Further filtering can be done, cf. Martin et al. (2022), but despite this, we find that the SweClarin resources are robust and provide results that can be used in our analyses.

4 Analyses and reflections on usability

Content analysis on the 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 and a topic analysis was carried out.

To assess the meaning that organisations ascribe to preventive innovations along the five attributes (as suggested by Rogers (1995), see Cahpter 2), 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.

4.1 Word clouds

Word clouds were employed as an initial step in our research to swiftly identify and visualise the most frequent words in the corpus generated through web crawling. The word clouds were created from the unprocessed scraped data, including pages unrelated to ISO/IEC 27001, accompanied by raw frequency data for each adoption type, to a holistic understanding of the content and gain insights into the overall discourse, beyond just their adherence to the ISO/IEC 27001 standard.

To build the word clouds we used the WordCloud package[7]. After filtering out stop words and a variety of other text strings, such as URLs and numbers, we used the 1000 most common words to build word

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

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

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

clouds for all companies. We also built word clouds, from the same data, for each of the preventive innovation adoption types. Figure 2[8] depicts such word clouds from two different adoption types clearly showing that there are differences between the two types as well as similarities, e.g. *service, security* and *information* are important in both but also that, for instance, *risk* is more prevalent for type 22.

Figure 2: English word cloud examples, adoption type 21 left and 22 right

Word clouds were particularly beneficial for providing the management researchers with an accessible overview of the most frequent words in the corporate communication of Swedish companies about ISO27001. The visualisation of key terms and their prominence assisted in pinpointing the primary focus areas within the corpus. The word clouds served as a foundational tool, offering a snapshot of the language and terminology prevalenty used in the industry's conversations about ISO27001.

Their primary role was to orient the research process by providing a basic overview of the corpus. However, word clouds offer limited analytical depth, primarily showcasing word frequency without delving into the nuances of context, relationships, or patterns among the terms. As such, they were more of a stepping stone towards more sophisticated analysis methods rather than a source of conclusive insights. Therefore, while instrumental in the early stages of research, word clouds were ultimately superseded by more advanced analytical tools that offered deeper, more contextually rich insights suitable for the final presentation of our findings.

4.2 Topic analysis

Following the initial analysis with word clouds, topic modelling was introduced as a method to further dissect and understand the corpus. For topic modelling we use the Gensim implementation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

Topic modelling is done both on the whole text and on the paragraphs containing an ISO27001 related term, as presented in Section 3. Further filtering is done by applying a frequency threshold and a threshold for spread. For LDA each paragraph is treated as a document and we use the Sparv parsed version to only include content words, in our case words with one of the part-of-speech tags noun, adjective, verb and adverb. We trained multiple topic models with different hyperparameters and used the NPMI coherence measure (Röder et al., 2015) to find optimal topics for each adoption type. Finally we use pyLDAvis[9] to visualise the topics, see Figure 3.

We also experimented with interpreting the results from the topic analysis using ChatGPT-3.5. We used a very simple prompt, *Interprete the following topic model we internally name "indirect economic impact, grouped 112131"* and the raw data from the topic analysis, see Figure 4 which shows data from four of the twelve most frequent terms in Figure 3. The management researchers considered the results from ChatGPT more useful than the visualisations, see Figure 5.

This technique enabled us to extract abstract topics, revealing hidden thematic structures within the corporate communications of Swedish organisations on ISO27001. Topic modelling provided a more

[8]The examples in Figure 2 are from the English corpus. Used here only for illustration as we do not normally analyse the English texts.

[9]https://pypi.org/project/pyLDAvis/

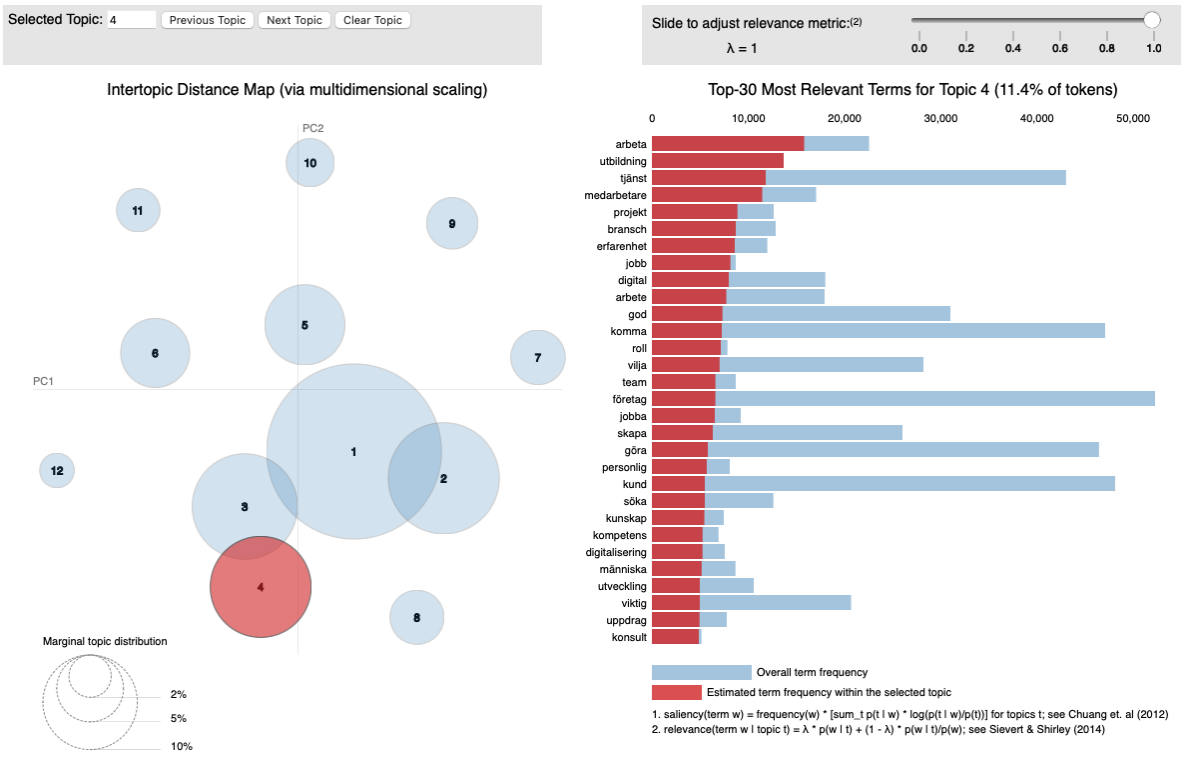

Figure 3: LDAvis visualisation of adoption type 11 for topic 4 (out of 12 topics as determined from the NPMI coherence measure). Translation of the terms from top to bottom: to work, education, service, coworker, project, sector, experience, job, digital, work, good, arrive/come, role, want, team, company, work, create, do, personal, customer, search, knowledge, competence, digitisation, human, development, important, mission and consultant.

```
    Interprete the following topic model we internally name "indirect economic
impact, grouped 112131":
LdaModel(num_terms=29019, num_topics=12, decay=0.5, chunksize=100)
[(0, '0.021*"arbeta" + 0.018*"utbildning" + 0.015*"tjänst" + 0.015*"medarbetare"
' '+ 0.012*"projekt" + 0.011*"bransch" + 0.011*"erfarenhet" + 0.011*"jobb" + '
'0.010*"digital" + 0.010*"arbete"'),
(1, '0.027*"ny" + 0.026*"år" + 0.017*"stor" + 0.017*"komma" + 0.016*"företag"
+ ' '0.015*"svensk" + 0.013*"kommun" + 0.009*"samarbete" + 0.008*"säga" + '
'0.008*"ta"'),
(2, '0.141*"kontakta" + 0.113*"följa" + 0.092*"meny" + 0.061*"söka" + '
'0.046*"beskrivning" + 0.041*"produkt" + 0.026*"ledig" + 0.024*"menya" + '
'0.023*"integritetspolicy" + 0.021*"recension"'),
(3, '0.091*"läsa" + 0.066*"språk" + 0.049*"tjänst" + 0.047*"innehåll" + '
'0.045*"välja" + 0.043*"kontakt" + 0.036*"sida" + 0.025*"nyhetsbrev" + '
'0.019*"mycken" + 0.018*"text"'),
............
```

Figure 4: ChatGPT prompt

nuanced understanding of the corpus, uncovering the prevalent themes and areas of interest in the virtual discourse of these companies. It was especially effective in identifying commonalities across various industries, illustrating the overarching trends and concerns within the sector. This level of analysis was vital in exploring different data governance approaches – agency, stewardship, brokerage – as well as the direct and indirect economic impacts of ISO27001 on these organisations. Topic modelling thus served as a critical tool in deepening our understanding of the corpus and shaping the theoretical framework for our research. Like word clouds, topic modelling was invaluable in guiding our research trajectory and

```
Based on the provided topic model "indirect economic impact, grouped 112131," here
is an interpretation of the topics:

Topic 0: This topic seems to be related to work and education, with terms such
as "arbeta" (work), "utbildning" (education), "tjänst" (service), and "projekt"
(project) being prominent. It might indicate discussions about skills, experience,
and digital work.

Topic 1: This topic revolves around new developments and companies, including terms
like "ny" (new), "år" (year), "företag" (company), and "kommun" (municipality). It
suggests discussions about new ventures, collaborations, and possibly local economic
impact.

Topic 2: This topic focuses on contact information and following updates. It
includes terms like "kontakta" (contact), "följa" (follow), and "integritetspolicy"
(privacy policy). It might pertain to discussions about communication channels,
privacy regulations, and staying informed.

Topic 3: This topic relates to content consumption and selection, with terms like
"läsa" (read), "språk" (language), "innehåll" (content), and "välja" (choose) being
significant. It suggests discussions about reading materials, multilingual content,
and selecting relevant information.

............
```

Figure 5: ChatGPT result. Translations included in the original response from ChatGPT.

informing our analytical approach.

4.3 Sentiment analysis

To get a more targeted investigation into the attitudes and perceptions surrounding ISO27001 we used sentiment analysis along the five attributes: relative advantage, compatibility, complexity, trialability and observability, see Chapter 2. To capture various uses of the attributes, synonyms were generated for each attribute by using the Gensim package[10] (Řehůřek & 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 eight native Swedish speaking innovation and entrepreneurship researchers at Linköping University to compile a final set of synonyms for the five attributes.

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).

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 assigns different sentiment values to different senses of a word, for instance, the Swedish word *fara* can mean 'danger' or 'go (away)' where the former has a negative sentiment and the latter is neutral. 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.

Vader weights the sentiment and gives a sentiment score between -1 and 1, where $\leq$ -0,05 represents a negative sentiment, a score between $>$ -0,05 and $<$ 0,05 signifies a neutral sentiment and, a score $\geq$

[10] https://radimrehurek.com/gensim/

0,05 denotes a positive sentiment to each attribute. The higher the value the more positive the attitude, and the lower the value the more negative the attitude regarding a specific attribute.

Vader also uses booster words, such as *scarcely* (Swedish *knappast*), to further refine the sentiment analysis. The booster dictionary used in these analyses is an 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).

The mean sentiment score for each of the Roger's attributes is calculated from the sentences in paragraphs with ISO sentences containing words, and synonyms, related to the attribute, as presented above. All words generated as synonyms to the Roger's concepts, see above, are lemmatized using the Stanza pipeline (Qi et al., 2020)[11] for Swedish and can easily be matched to their occurrence in the text as the corpus is parsed using the Sparv pipeline, which includes the Stanza pipeline for Swedish lemmatization as one lemmatizer. Typical examples from the analyses can be seen in Figure 6[12].

Results from the sentiment analysis for each of the six adoption types along the five Rogers' attributes is presented in Table 3.

Table 3: Mean sentiment score for each adoption type and Roger's attribute

	11	12	21	22	31	32
Relative Advantage	0.172	0.111	0.128	0.174	0.000	0.092
Compatability	0.124	0.046	0.129	0.033	0.125	0.009
Complexity	0.156	0.135	0.110	0.077	0.122	0.091
Trialability	0.150	0.128	0.124	0.119	0.115	0.093
Observability	0.141	0.096	0.124	0.073	0.114	0.091

As can be seen in Table 3 the scores indicate that the texts are rather neutral, with adoption type 32 being more neutral than the other. Pairwise Welch t-tests for each category compared to the category's overall score also shows that this difference is significant for all Roger's attributes[13], i.e. paragraphs with ISO sentences comprising words related to any of the Roger's attributes for Broker companies that have direct economic benefits from communication adoption, adoption type 32, are more neutral.

The neutral sentiment example with Roger's attribute Relative Advantage, Table 3 is on the one hand a typical example of a neutral paragraph; it expresses no sentiment. But it also illustrates the nature of the corpus, in this case it is merely a number of statements or phrases, probably taken from a table of content or a menu on a web page. Nevertheless it has been parsed by the Sparv pipeline, showing the robustness of the SweCLARIN resources.

Table 3 further shows that no mean sentiment is negative. This is not surprising as the corpus contains companies web pages, pages where the companies present themselves.

However, as can be seen from Figure 6 even negative paragraphs are sometimes to be regarded as positive, i.e. the example from adoption type 32. That example also shows a problem with lexical sentiment analyses. When sentiment is based on averaging over words that have a positive or negative sentiment without taking context into account. Then sometimes phrases such as *increases the possibilities of identifying* in the example from adoption type 32 in Figure 6 that modifies the negative words *threats, risks, lost, stolen* are missed and the phrase is classified as negative when in fact is should be positive.

As this could potentially affect our results we have investigated the distribution of sentences with positive (≥ 0.05), negative (≤ -0.05), or neutral sentiment (between -0.05 and +0.05), as proposed by Hutton and Gilbert (2014), see Table 4.

As can be seen in Table4 the amount of sentences with positive, negative or neutral sentiment is equally distributed among the various adoption types. In each adoption type, the proportion of negative

[11] https://stanfordnlp.github.io/stanza/

[12] Translated by google translate

[13] $p < .001$ for all pairs.

Adoption type: 21
Roger's attribute: Trialability
Mean sentiment score: 0.612
Paragraph: Our NMT solution is developed in-house by our machine learning team of data scientists and engineers. Our neural machine translation is ISO 27001 compliant and all information is stored in our private data cente, so you can trust it to be secure.
Swedish original: *Vår NMT-lösning har utvecklats in-house av vårt maskininlärningsteam med datavetare och tekniker. Vår neurala maskinöversättning följer ISO 27001 och all information lagras i vårt privata datacenter, så att du kan lita på att den är säker.*

Adoption type: 32
Roger's attribute: Relative Advantage
Mean sentiment score: 0.0
Paragraph: Introduction to information security Different types of information security threats The requirements of ISO 27001 and interpretation of the requirements Exercises in interpreting and applying the standard in practice Good examples Culture and commitment in the workplace The red thread of the management system Tools for improvement Integration of systems
Swedish original: *Introduktion till informationssäkerhet Olika typer av informationssäkerhetshot Kraven i ISO 27001 och tolkning av kraven Övningar i att tolka och att tillämpa standarden i praktiken Goda exempel Kultur och engagemang på arbetsplatsen Ledningssystemets röda tråd Verktyg för förbättring Integrering av system*

Adoption typ: 32
Roger's attribute: Trialability
Mean sentiment score: -0.25
Paragraph: A certification according to ISO 27001:2017 ensures that you work in a systematic and effective way with information security. With the support of the standard you can create a framework for how you protect your most important information. Following a standard in your work with information security increases the possibilities of identifying threats and preventing risks of information being lost or stolen.
Swedish original: *Ett certifiering enligt ISO 27001:2017 säkerställer att ni arbetar på ett systematiskt och effektivt sätt med informationssäkerhet. Med stöd av standarden kan ni skapa ett ramverk för hur ni skyddar er viktigaste information. Att följa en standard i ert arbete med informationssäkerhet ökar möjligheterna att identifiera hot och förebygga risker för att information försvinner eller stjäls.*

Figure 6: Examples from the sentiment analyses

Table 4: Number of sentences with positive, negative or neutral sentiment for each adoption type

	11	12	21	22	31	32
POSITIVE	81,383	4,536	52,337	10,877	8,692	32,276
NEGATIVE	10,371	894	7,085	3,161	1,337	10,138
NEUTRAL	95,127	5,249	60,772	13,624	9,217	35,068

sentiment, relative to the total of positive and negative sentiments, ranges from 5-15%. Adoption type 11 exhibits the lowest proportion, with approximately 6% of sentences expressing negative sentiment in its entire corpus. In contrast, adoption type 32 has the highest, at around 15%. This indicates that companies generally present themselves in a positive light in their corporate presentations. We, thus, assume that the problem with sentences' sentiment is the same for all types, meaning that we can compare the sentiment of various adoption types but not necessarily give an absolute sentiment score.

5 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 cyber security risks as one of the great challenges of our time. Using web scraping tools and a variety of computational linguistics tools, we were able to extract and analyse large amounts of text. These texts on preventive innovation ISO/IEC 27001 include communicative efforts published on 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) need 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 management researchers and the SweClarin language technology researchers has been imperative for the success of this project. Based on the needs of the management researchers various analyses have been performed, and assessed. It was, for instance, initially assumed to be important to use word clouds to give an illustration of word frequencies amongst the various adoption types. Word clouds were good to get a coarse understanding of concepts used in the corpus and guided the management researchers' further investigations. Topic models were further used to guide the deep readings. However, they were also rather diverse and hard to interpret and did not form a clear characteristics of the various adoption types. The use of LLMs, in this case ChatGPT, turned out to be an interesting, and useful, complement to the classical visualisations of topic models.

In further discussions with the management 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 management researchers the opportunity to see a quantification of the meanings to further aid the deep readings. These insights gave rise to new perspectives, which resulted in analyses along different dimensions of adoption type. Future research on sentiment analysis with the management researchers include looking into the actual sentiment analyses using new models. The Swedish BERT model (Hägglöf, 2023)[14] is one such possibility that needs to be assessed and compared to the lexical approach using the SweCLARIN SenSALDO 0.2 sentiment lexicon.

[14]https://huggingface.co/KBLab/robust-swedish-sentiment-multiclass

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*.

Bergquist, W. (1993). *The postmodern organization: Mastering the art of irreversible change*. San Franscico: Jossey-Bass.

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*.

Downs, J., W., G., & Mohr, L. B. (1976). Conceptual issues in the study of innovation. *Administrative Science Quarterly!*, *21*(4), 700–714.

Fineberg, H. (2013). The paradox of disease prevention: Celebrated in principle, resisted in practice. *JAMA*, *310*(1), 85–90. https://doi.org/doi:10.1001/jama.2013.7518

Fried, A., Jönsson, A., Dragisic, S. P., & Bandyopadhyay, S. (2024). Preventive innovation revisited: Towards a sitational and meaning-based understanding of rogers' communication concept [Paper not published]. *84. Wissenschaftliche Jahrestagung des Verbandes der Hochschullehrerinnen und Hochschullehrer für Betriebswirtschaft 2024*.

Geissinger, A., Laurell, C., Öberg, C., & Sandström, C. (2023). Social media analytics for innovation management research: A systematic literature review and future research agenda. *Technovation*, *123*(102712).

Hägglöf, H. (2023). The kblab blog: A robust, multi-label sentiment classifier for swedish. https://kb-labb.github.io/posts/2023-06-16-a-robust-multi-label-sentiment-classifier-for-swedish/

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

Lo, L. K., & Chang, D. S. (2007). The difference in the perceived benefits between firms that maintain iso certification and those that do not. *International Journal of Production Research*, *45*(8), 1881–1897.

Martin, L., Fan, A., de la Clergerie, É., Bordes, A., & Sagot, B. (2022, June). MUSS: Multilingual unsupervised sentence simplification by mining paraphrases. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 1651–1664). European Language Resources Association. https://aclanthology.org/2022.lrec-1.176

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.

Orlikowski, W., & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. *Administrative Science Quarterly*, *39*, 541–574.

Putnam, L. L. (1999). Shifting metaphors of organizational communication: The rise of discourse perspectives. In P. Salem (Ed.), *Organizational communication and change* (pp. 45–65). Cresshill, NJ: Hampton Press.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

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.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*, 399–408. https://doi.org/10.1145/2684822.2685324

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.

Saura, J. R., Palacios-Marqués, D., & Ribeiro-Soriano, D. (2023). Exploring the boundaries of open innovation: Evidence from social media mining. *Technovation*, *119*(102447).

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

Taylor, J. R. (1993). *Rethinking the theory of organizational communication: How to read an organization*. Norwood, NJ: Ablex.

Tornatzky, L. G., & Klein, K. J. (1982). Innovation characteristics and innovation adoption-implementation: A meta-analysis of the findings. *IEEE Transactions on Engineering Management*, *EM-29*, 28–43.

Yates, J., & Orlikowski, W. (1992). Genres of organizational communication: A structurational approach to studying communication and media. *Academy of Management Review*, *12*(2), 299–326.