Master’s Programme in Life Science Technologies

Sentence Embeddings in Topic Modeling

Analysing News Articles through Cybersecurity Concepts

Lauri Pykälä

Aalto University

MASTER'S THESIS

Aalto University
MASTER'S THESIS 2022

Sentence Embeddings in Topic Modeling

Analysing News Articles through Cybersecurity Concepts

Lauri Pykälä

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.
November 29, 2022

Supervisor: Prof. Jouko Lampinen
Advisors: MSc Tuomas Takko, Prof. Kimmo Kaski

Aalto University
School of Science
Master's Programme in Life Science Technologies

Aalto University

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Abstract

Author
Lauri Pykälä

Title
Sentence Embeddings in Topic Modeling: Analysing News Articles through Cybersecurity Concepts

School School of Science

Master's programme Life Science Technologies

Major Complex Systems **Code** SCI3060

Supervisor Prof. Jouko Lampinen

Advisors MSc Tuomas Takko, Prof. Kimmo Kaski

Level Master's thesis **Date** November 29, 2022 **Pages** 59 **Language** English

Abstract

Emerging cybersecurity threats warrant a rethinking of effective countermeasures at the strategic level. Treating technology as isolated from geopolitics carries the risk of neglecting the vital role that mutual trust sustains in the security of critical infrastructure like 5G networks. National cybersecurity strategies serve a dual purpose. Primarily, they are guidelines for tackling practical concerns. Moreover, they can inform the public about the necessity of the compromise between liberty and security.

Mass media have a vital role as the 'fourth branch of government' in framing the public debate around potentially controversial policies. Hundreds of articles published between 2011 and 2020 by the Finnish public service media company include mentions of Huawei, a Chinese telecommunications company entangled in a debate of cybersecurity, geopolitics and technologies.

Computational text analysis is a toolkit to facilitate qualitative research on corpuses that are too large for manual analysis. One widely adopted approach, known as topic modelling, is based on analysing statistical patterns of co-occurring words to obtain higher-level themes.

Modern language models are trained on enormous volumes of textual data to become capable of 'filling in the blanks' for new sentences. As an intermediate step, numerical representations of text called embeddings are produced. Repurposing sentence embeddings from such models is being investigated as an alternative approach to topic modeling.

In this thesis, embeddings from a BERT type language model were computed for all sentences in the Yle news dataset that mentioned Huawei. The 768-dimensional sentence vectors were projected into lower-dimensional representations using UMAP. A graphical interface was developed to explore the 2D vectors as scatter plots, and the corresponding 5D ones were clustered with HDBSCAN. This visualisation was interactively probed to fine-tune the collection of parameters in order to obtain interpretable groupings of sentences. Keyword lists were compiled for each cluster using a version of TF-IDF to describe each sentence clusters' characteristic nouns, verbs and adjectives.

A workflow based on thematic analysis was developed to refine the computational results into conceptual abstractions. The cluster keywords were categorised, based on which three general themes were then found. Themes included allegations of misconduct levied at the company and their responses to them, discussion of the telecommunications market as well as coverage of products from a consumer perspective. The significance of these themes is discussed in relation to the background concepts from cybersecurity strategies.

Keywords topic model, sentence embedding, natural language processing, cybersecurity

A! Aalto-yliopisto

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tiivistelmä

Tekijä
Lauri Pykälä

Työn nimi
Lauseupotukset aihemalleissa: Uutisartikkelien analysointi kyberturvallisuuskäsitteiden kautta

Korkeakoulu Perustieteiden korkeakoulu

Maisteriohjelma Life Science Technologies

Pääaine Complex Systems **Koodi** SCI3060

Valvoja Prof. Jouko Lampinen

Ohjaajat DI Tuomas Takko, Prof. Kimmo Kaski

Työn laji Diplomityö **Päiväys** 29. marraskuuta 2022 **Sivuja** 59 **Kieli** englanti

Tiivistelmä

Uudenlaiset kyberturvallisuuden uhkakuvat edellyttävät tehokkaita vastatoimia, strategiselta tasolta alkaen. Teknisiin ratkaisuihin keskittyminen erillään geopoliittisista riskeistä sivuuttaa kauppakumppanien välisen luottamuksen merkityksen kriittisen infrastruktuurin, kuten 5G-verkkojen turvaajana. Kansallisilla kyberturvallisuusstrategioilla on kaksoisrooli. Ensisijaisesti ne toimivat käytännön toimien ohjaajina. Tämän lisäksi strategiat toimivat suuren yleisön informointitarkoituksessa kysymyksissä, jotka liittyvät vapauden ja turvallisuuden välisten kompromissien hakemiseen.

Joukkoviestimillä on tärkeä tehtävä ‘neljäntenä valtiomahtina’ julkisen keskustelun kehystämisessä kiistanalaisten kysymysten ympärillä. Sadat Yleisradion vuosina 2011-2020 julkaisemat artikkelit sisältävät viittauksia Huaweihin. Kyseinen kiinalainen tietoliikennealan yritys on kyberturvallisuuden, geopolitiikan ja teknologian keskustelun keskiössä.

Laskennallinen tekstianalyysi on työkalupakki, joka mahdollistaa määrällisen tutkimuksen suurille aineistoille, joiden manuaalinen analysointi on poissuljettua. Eräs yleisesti käytetty menetelmä, aihemallinnus, perustuu yhdessä esiintyvien sanojen tilastollisten lainalaisuuksien laskentaan ja ylemmän tason teemojen koostamiseen.

Uusimmat kielimallit koulutetaan valtavilla tekstiaineistoilla, tehtävänään ‘täyttää puuttuva kohta’ myös aineiston ulkopuoliselle materiaalille. Välivaiheenaan mallit tuottavat teksteistä numeerisia vektoreita eli upotteita. Näiden upotteiden hyödyntämistä on tutkittu vaihtoehtoisena lähestymistapana aihemallinnukseen.

Yleisradio-aineistoon sisältyneet Huaweihin liittyvät uutistekstit syötettiin BERT-tyyppiseen kielimalliin. Näin saadut 768-ulotteiset lauseupotusvektorit projisoitiin pienempiulotteisiksi UMAP:lla ja klusteroitiin HDBSCAN:lla. Työssä kehitettiin myös graafinen käyttöliittymä lauseupotusten tutkimiseen kaksiulotteisina sirotekuvaajina. Visualisaation avulla hienosäädettiin algoritmien säätöarvoja tulkittavien lauseryhmittelyjen muodostamiseksi. Lauseryhmille laskettiin avainsanalistat sanaluokittain TF-IDF -menetelmällä.

Lisäksi kehitettiin temaattiseen analyysiin perustuva työnkulku laskennallisten tulosten jalostamiseksi ylemmän tason teemoiksi. Lauseryhmistä löytyi kolme yleisteemaa. Teemoihin kuuluivat Huaweita kohtaan esitetyt väärinkäytössyytökset sekä yrityksen reaktiot niihin, yleinen keskustelu tietoliikennemarkkinoiden kilpailutilanteesta sekä matkaviestintuotteiden kuluttajalähtöinen esittely. Näitä teemoja analysoitiin suhteessa kansallisten kyberturvallisuusstrategioiden avainkäsitteisiin.

Avainsanat aihemalli, lauseupotus, luonnollisen kielen käsittely, kyberturvallisuus

Acknowledgements

I would like to express my gratitude for the support, insights and feedback provided by Kimmo, Tuomas and Kunal as well as Pertti & others at Cyberwatch Oy!

Otaniemi, 28.11.2022 Lauri Pykälä

Table of Contents

List of Tables

List of Figures

Abbreviations

BERT Bidirectional Encoder Representations from Transformers. 3, 11, 12, 16–18, 27, 41, 42

eom excess of mass. 21

GPT-3 Generative Pre-trained Transformer 3. 43

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise. viii, 27, 28, 39

LDA Latent Dirichlet Allocation. 13, 14, 18, 41

MLM Masked Language Model. 16

MST minimum spanning tree. 22

NCSS National Cybersecurity Strategy. 4, 8

NER Named Entity Recognition. 12, 24

NLP Natural Language Processing. 2, 3, 10, 14, 17

NSP Next Sentence Prediction. 16

OOV Out of Vocabulary. 11, 15

PCA Principal component analysis. 19

PDF Probability density function. 20, 21

PLM Pretrained language model. 41

PoS Part of Speech. 11, 24

TF-IDF Term Frequency-Inverse Document Frequency. 18, 24

t-SNE t-distributed stochastic neighbor embedding. 19

UMAP Uniform Manifold Approximation and Projection. viii, 19, 20, 27

1. Introduction

In 2022, an online game called Semantle[1] was released, and by April it was accessed by 200 000 daily players. It made clever use of a semantic similarity algorithm to give feedback on guesses, with the goal of finding the day's 'secret word'. [1]

The measure of similarity in the game is based on converting each word in a vast corpus of news articles into a high-dimensional vector, aiming to place words that occur in the same contexts in the articles close to each other in the vector space. The similarity scores are computed based on distance to the 'secret word'. To get an idea of what this entails, a walkthrough of an example game is presented in Fig. 1.1 and Fig. 1.2.

Note that the algorithm powering the game, called 'word2vec' (further discussed in Section 4.3), was published already in 2013 [2] and falls far short in performance on most evaluation benchmarks when compared to the current state-of-the-art methods. Nevertheless, the similarity metric derived from it is capable of conveying diverse dimensions of meaning interpretable enough to solve the game. Even in the context of the game, though, these results only hold for a narrow subset of the semantic categories or perhaps only when the similarity score is calculated in relation to a specific word.

Some implications stemming from the interactive, 'human-in-the-loop' nature of Semantle should be considered before basing any general claims on what word2vec is capable of. For instance, the scores are only displayed for words entered as guesses, effectively biasing the results towards a human-informed shortlist of candidates. Playing the game, I developed a degree of intuition for the games's representation of semantic similarity, even before being familiar with its underlying theory. Online commentators have similarly noted that 'what makes this challenging is that the

[1]Available at `https://semantle.novalis.org/`

After some early random guesses on the initial hints, I noticed that the word ‘oil’ had the highest score, and was thus the best guess. Thinking that the secret word might be related to the energy industry, I tried to confirm this. When ‘fuel’, ‘money’ and ‘trade’ did not improve the score, I had to discard the hypothesis. Perhaps another type of ‘oil’ was a better direction? This produced a breakthrough; food-related guesses (‘potato’, ‘pizza’) appeared to get me closer to the secret word, and the word ‘water’ was the best candidate so far. The next clue emerged from beverage-related words; for some reason, ‘wine’ and ‘juice’ had lower scores than ‘cola’ and ‘water’. The answer to this algorithmic puzzle turned out to be that the latter are more commonly discussed in a context also mentioning ‘ice’. Using ‘ice’ as a new reference point quickly yielded results; after a total of 70 guesses, the secret word turned out to be ‘melting’.

Figure 1.1. An example playthrough of Semantle.

#	Guess	Similarity	Getting close?
70	melting	100.00	FOUND!
67	glacier	44.32	975/1000
57	ice	40.91	938/1000
53	freezing	39.33	898/1000
69	temperature	39.26	891/1000
56	icy	37.19	826/1000
54	frozen	36.75	799/1000
59	snow	32.96	473/1000
52	cold	29.87	(tepid)
58	iced	29.10	(tepid)
26	glass	21.05	(tepid)
20	water	20.37	(tepid)
55	freezer	19.68	(cold)
23	cola	19.60	(cold)

Figure 1.2. At each turn, the player types a guess for the secret word and receives feedback on how similar the guess and the secret word are. Screenshot from `semantle.com`.

similarity number is boiling down potentially hundreds of different axes by which a word might be considered “similar”’ [3].

The example of this simple online game invoked some motivating questions pertinent to the themes explored in this thesis, including: How do computationally derived metrics of similarity align with or deviate from human intuitions? What role does interactivity play in the usability of Natural Language Processing (NLP) systems? Finally, how could shortcomings of language algorithms be mitigated with suitable design choices for interfaces?

1.1 Preamble

NLP research in computer science publications typically consists of well-established steps; first presenting a novel idea or improving on previous work, then evaluating the performance on a manually created dataset of 'ground truth' annotations, and finally benchmarking the method against existing baselines. In qualitative research, on the other hand, the entire notion of a 'ground truth' is elusive, even inapplicable. Research workflows are often ad hoc and unstandardised regarding how the computational aspects fit into the overall process.

Nguyen, Liakata, DeDeo, *et al.* [4] discuss the toolkit of computational text analysis, an interdisciplinary field characterised by the integration of methods from different research traditions. They describe a general framework and propose some guiding questions. These provide the framework based on which this thesis is structured into chapters.

The main steps are identifying and formulating research questions, their conceptualisation, choosing the dataset and preprocessing methods, operationalisation and analysis.

In Chapter 2, the focus is on *what* this research is on. To this end, background on national cybersecurity issues is presented first, followed by a section arguing that debate on 5G network vendors is not just a question of technology. The chapter concludes by introducing the case of Huawei.

Chapter 3 argues that news media is essential in informing the public on cybersecurity issues and thus a suitable resource of study. The Yle news dataset is introduced, including the filtering and preprocessing steps.

In Chapter 4, we shift to *how* the research questions will be answered. A brief history of topic models is presented, including some theory and practical applications. This is followed by outlining a different, more recent paradigm of computational text analysis: the deep-learning-based pre-trained language models, specifically one called Bidirectional Encoder Representations from Transformers (BERT).

Chapter 5 describes implementation of the methodology, consisting of the unsupervised computational analysis, interacting with the model to explore the data and the qualitative thematic analysis undertaken.

The computational and qualitative analysis results are presented in Chapter 6 and related to the background concepts. Finally, in Chapter 7, new insights relevant to the research questions are discussed, the methodology is evaluated, and some leads for future studies are suggested.

2. Conceptualisation of cybersecurity issues

Conceptualisation is the task of converting research questions to high-level abstractions, which will inform the interpretation of results. The background concepts discussed in this chapter are informed by features of National Cybersecurity Strategies (NCSS), with particular attention to the questions related to critical infrastructure such as 5G networks.

2.1 National Cybersecurity Strategies as a theoretical framework

National Cybersecurity Strategies are policy documents that, according to European Union Agency for Cybersecurity (ENISA), outline a ‘plan of actions designed to improve the security and resilience of national infrastructures and services’.

Limnéll *et al.* note that commonalities exist in National Cybersecurity Strategy (NCSS)s from different countries, including aspects of vulnerabilities and suggestions for countering them. In addition to serving national interests, NCSS can be helpful guidelines for private companies.

Limnéll *et al.* categorise typical features of national cybersecurity strategies into seven elements, summarised in Fig. 2.1. This serves as the conceptual background to provide perspectives through which the case study of the security concerns surrounding Huawei and their impacts can be viewed.

NCSS commonly assert that fostering cooperation and communication across administrative branches, private companies, and citizen organisations is essential. Particular attention should be given to ensuring the timely sharing of accurate threat intelligence, to maintaining good situational awareness and looking ahead in order to operate proactively, as opposed to reactively. This should be tasked to a dedicated central authority, i.e. a cybersecurity centre.

International cooperation should be encouraged; global threats warrant global solutions and cooperation. Exchanging know-how through research and development partnerships should be strengthened.

Legislation, both domestic laws and international treaties, should be reviewed to ensure they are fit for purpose in facilitating the mounting of responses. The primary responsibility of national governments is shielding critical infrastructure, both physical and digital.

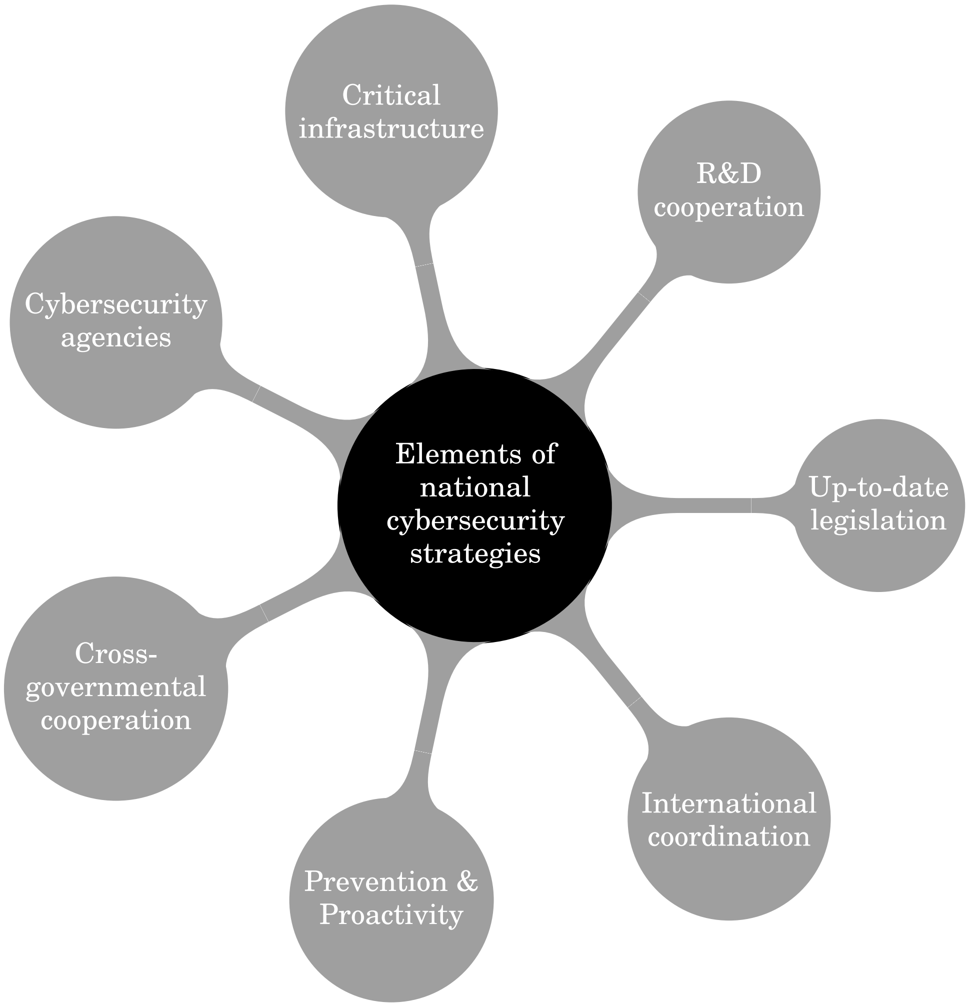


Figure 2.1. Themes found in national cybersecurity strategies by Limnéll, Majewski, and Salminen [6]

2.2 Securing 5G infrasctucture is a strategic issue

The rollout of 5G, the next generation of mobile networks, offers a textbook example of critical infrastructure. 5G's promises include greater speeds, lower latency and explosive growth in the total number of devices connected to the Internet. [7]

In addition to the promise of currently unimaginable leaps in economic efficiency in automation and AI applications, the downsides are also evident. As the number of connected devices grows, so does the size of the attack surface vulnerable to being exploited, either as part of a botnet or enabling espionage on an unprecedented scale.

It should be acknowledged that the trade of technology, including weapons, between parties that do not trust each other, has always occurred. Risks have been mitigated by inspecting and auditing the wares sold, thereby shifting the burden of trust from the trading partner to one's own verification process. [7]

Keeping this in mind, it is fair to ask whether some technical screening procedure could alleviate security concerns around 5G infrastructure. For example, the conclusion reached by Huawei Cyber Security Evaluation Centre (HCSEC) Oversight Board [8] was that 'only limited assurance [can be provided] that all risks to UK national security [...] have been sufficiently mitigated'. Similarly, Lysne [9] lamented that the prospects for a robust trust-free equipment verification process are 'rather bleak'. In short, a technical inspection cannot sufficiently mitigate the risks.

The first challenge is that the amount of code in modern software systems is prohibitive to manual analysis, even if made available for inspection. Furthermore, the screening would have to be repeated each time the software updates.

Machine learning approaches like those used in antivirus software are useless when the potential malicious actor is the software vendor instead of a third party. This is because such methods rely on having a system known to be unaffected ('healthy') and comparing it to the one being investigated. Such a baseline is not available when looking for backdoors hiding in every copy of the software.

Malicious activity could also be detected 'as it happens' by inspecting network traffic. However, this approach is hindered when the signal of interest is small compared to the total amount of bandwidth, as is the case with 5G base stations.

One category of malicious activity is usually easier to detect, at least on a large scale: sabotage. Yet, being able to detect damage that has already taken place is of little use.

The conclusion that technical solutions are insufficient to tackle the security concerns should provoke a frame shift; '5G rollout needs to be recognised as a strategic rather than merely a technological choice.' [10].

2.3 Case Huawei

The primary impetus that transformed the discussion around 5G from a theoretical debate into a practical security concern can be traced to the ambitions of Chinese suppliers, particularly Huawei, to enter Western markets for 5G infrastructure.

Friis and Lysne [7] trace the earliest signs of the developments to warnings voiced in 2010 by the Federal Bureau of Investigation (FBI) and other experts that Chinese companies can threaten US security interests. Apart from a report by the US Congress [11] in 2012, the issue did not gain much attention in public discourse until 2016, when President Trump and his administration opted for an openly confrontational approach to China, both in rhetoric and practice. In 2019, Huawei was included on the US 'Entities List', barring them from using components including chips or software from American companies.

In Europe, a more lenient approach initially took hold. For example, in January 2020, the UK opted to cap Huawei base stations to a $35\,\%$ share while excluding only sites such as military bases. The US government reacted with strong opposition to the decision. Only six months later, the UK hardened their stance and announced the complete ban on installing new Huawei 5G equipment from 2021 and the removal of pre-existing ones by 2027. [12]

3. The Yle News dataset and preprocessing steps

This chapter opens by presenting the connection between ideas in cybersecurity strategies and news coverage. The Yle News dataset is introduced, followed by a description of the filtering and preprocessing steps applied.

Note that the designation of an analytical step as either preprocessing or part of the 'main' methodology is context-dependent. In this work, the computational steps using 'off-the-shelf' solutions are viewed as preprocessing. In contrast, the computational methods discussed in-depth and assessed in depth are presented in Chapter 4.

3.1 The role of media in cybersecurity

Jansson and Sihvonen [13] compare themes of cybersecurity discourses in Finnish news media and the Finnish NCSS. They note that the academic study of cybersecurity from a social science perspective is nascent, especially from a Finnish perspective.

A prominent feature of discourse around cybersecurity is securitisation, i.e. viewing issues through a perspective of threats to society. Significant incidents are often covered widely in the news, especially when a narrative can be constructed on the possible motives of the perpetrators. The abstract, non-physical nature of cyberattacks presents challenges for describing their impact. Since understanding the technical details and methods of an attack is usually beyond the interests or knowledge of a lay reader, stories often latch on quantitative features, such as the number of users affected or the estimated financial consequences, often compared to previous high-profile incidents. [13]

Similar to Limnéll *et al.*, Jansson and Sihvonen [13] argue that cybersecurity strategies have two main purposes and audiences. Their primary purpose does not differ from any other policy paper; an internal tool for official and private stakeholders to coordinate responses. The other aim is less apparent; cybersecurity strategies serve to communicate to citizens the actions taken to counter threats and justifying them.

Jansson and Sihvonen [13] argue that 'cybersecurity strategies are a means of dispelling the conflict-laden perception of states as surveillants of citizens' to make the case that actions taken are for the common good, not manifestations of an overreaching big brother. Optimally, strategy publications can 'evoke a sense of national unity' and even ideologically consolidate viewpoints.

Similarly to the recommendations on approaching 5G issues described in Chapter 2, journalistic outlets should be aware that 'cybersecurity is *not just about technology*, [and therefore] conversation on the topic should be part of a socially and politically alert debate on the state of the world and interstate relations.' [13, emphasis added]

3.2 The Yle News dataset

The dataset consists of all news articles published by the Finnish Broadcasting Corporation (Yle) in the years 2011-2020, split into two parts in the source material and totaling 832 488 articles. [1] [14], [15]

3.2.1 Filtering the articles

A subset of the data was selected, including all articles containing the string `huawe`, resulting in 317 articles and 11 145 sentences. Fig. 3.1 illustrates the number of times the search string appeared in the dataset. The first peak can be seen in late 2012, followed by a relatively stable number of mentions until a large increase in early 2019. In the original dataset, articles were divided into segments with varying lengths, corresponding to layout elements such as paragraphs, captions and subheadings. The texts were concatenated by article and then split at each period sign, to obtain individual sentences, which were once again limited to those containing a match for `huawei`. A histogram of the sentence lengths is seen in Fig. 3.2.

[1] [14] contains 703 673 articles published 2011-2018 and [15] contains 128 815 articles published 2019-2020. The datasets are available at `https://www.kielipankki.fi/corpora/`

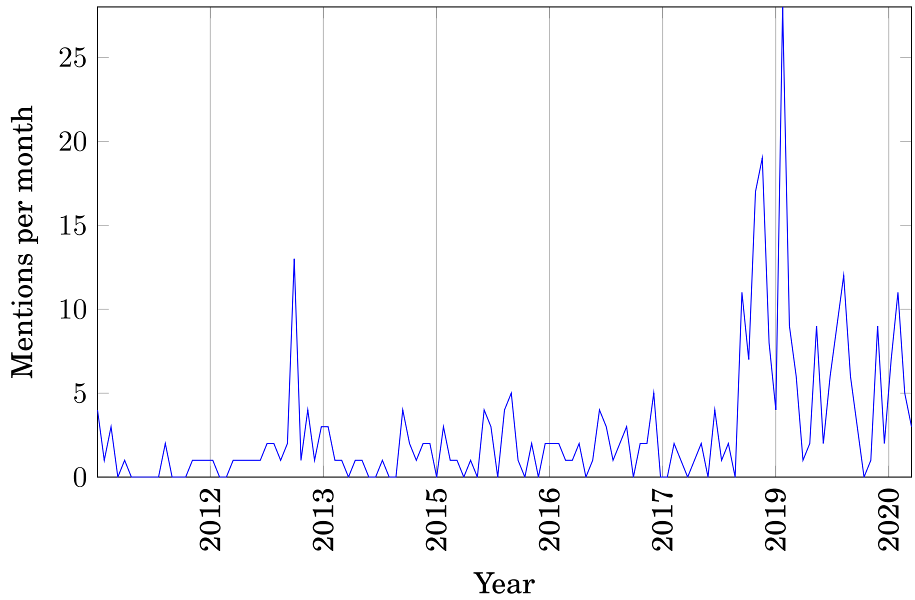


Figure 3.1. A time series showing the number of articles mentioning Huawei per month

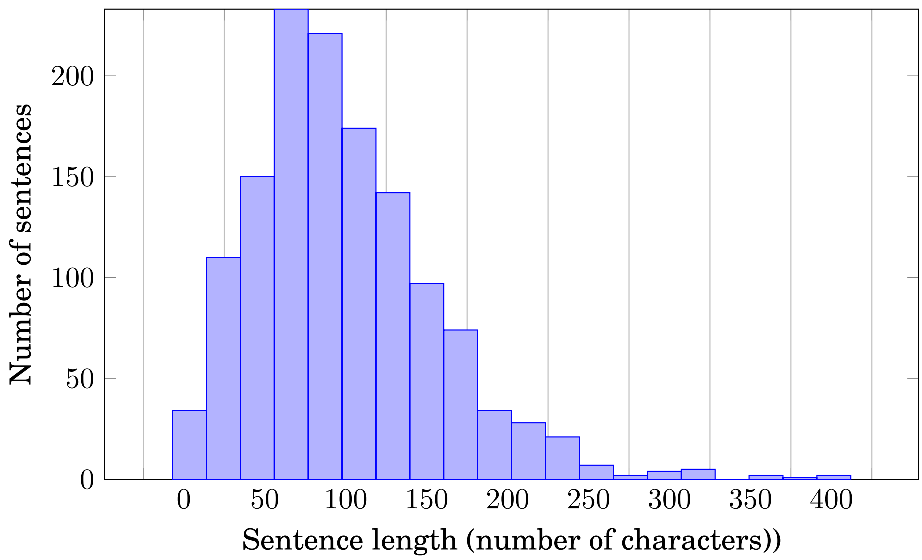


Figure 3.2. The median length of a sentence in the dataset is approximately 100 characters.

3.2.2 Preprocessing

Cleaning the text data consisted of removing URLs, certain Unicode special symbols and the Markdown syntax used in the online typesetting of the news (see Appendix B for details).

Tokenisation is a common NLP preprocessing step. In practice, the sentences' words are split to match the fixed-size vocabulary data of the language model using an algorithm called WordPiece. If the entire word is absent from the vocabulary, the tokenizer attempts to break the word into the longest constituent subwords found in the vocabulary or as a last resort, into individual characters. This allows retaining some information on the contextual meaning also for Out of Vocabulary (OOV) words, ie. those that the model has not 'seen' before.

An example tokenisation is shown in Table 3.1. There are several nested levels of indexing schemes in the text data after the preprocessing steps. The raw output from Bidirectional Encoder Representations from Transformers (BERT) includes a vector for each token. Words consist of one or more token, and similarly named entities can encompass more than one word. Consequently, the vectors for words are obtained from the means of their constituent tokens and the vectors for named entities by again computing the mean. Lemmatisation is the process of converting the in-

Label	Word	Token Index	Token	Token ID
NaN	NaN	0	[CLS]	102
O	Pääaineeni	1	Pää	1793
		2	##aine	41570
		3	##eni	1151
B-ORG	Aalto-yliopistossa	4	Aalto	9947
I-ORG		5	-	166
I-ORG		6	yliopistossa	9008
O	on	7	on	145
O	Complex	8	Comp	26581
		9	##lex	26513
O	Systems.	10	System	47534
		11	##s	50011
O	.	12	.	111
NaN	NaN	13	[SEP]	103

Table 3.1. Example of tokenization and NER labeling. Tokens, words and named entities have their own indexing schemes.

flected forms of words into their dictionary form, or lemma. Lemmatisation takes into account the part of speech and context of a given word instead of simply removing suffixes. Part of Speech (PoS) tagging is the process of assigning lexical categories to words in the corpus, such as nouns, verbs etc.

Tokenisation, lemmatisation, PoS tagging and dependency parsing tasks were performed using the Turku Neural Parser pipeline software by [16], based on the Turku Dependency Treebank (TDT), an annotated multidomain dataset of Finnish text [17], [18]. The reported accuracy (F1 scores) of the method reported in [19] are $96.66\,\%$ for PoS tagging, $95.32\,\%$ for lemmatisation and $99.69\,\%$ for tokenisation.

Lemmatisation was used to enhance the computation of cluster keywords by treating the various conjugated instances of a single dictionary word as one, and part-of-speech tagging further improved the interpretability and steerability of the keywords. The computation of cluster keywords is detailed in Section 4.7.3

While dependency parsing information was also obtained (see Fig. 3.3), its implementation as part of this topic modeling framework remains as a future direction of research. Named Entity Recognition was performed

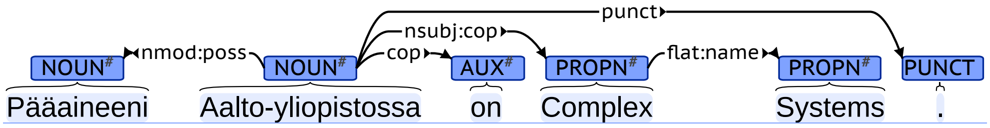


Figure 3.3. The arrows represent linguistic dependencies and the labels above words are part-of-speech tags

using the method and code by Luoma, Oinonen, Pyykönen, *et al.* [20]. It was trained using a corpus containing over $10\,000$ named entities from 754 Finnish texts covering different genres. Their evaluations showed that a BERT model trained on the annotations achieved precision and recall scores of over $90\,\%$ on the Named Entity Recognition (NER) task.

4. Operationalisation of the topic modeling framework

In operationalisation, the goal is to develop measures for the concepts of interest, namely the cybersecurity themes described in Section 2.1. This chapter begins with a brief history of topic modeling, with the widely used Latent Dirichlet Allocation (LDA) as the prototypical example. Pretrained language models, an entirely different family of approaches to computationally representing language, are then introduced. This is followed by examining the intersection of these lines of inquiry and the nascent research area of augmenting or even wholly replacing traditional topic models with more advanced methods.

4.1 Probabilistic topic modeling

The term *topic model* was coined by Blei and Andrew Y. Ng to describe their proposed Latent Dirichlet Allocation (LDA) algorithm. Topic models provide an automated process for coding large collections of texts into a set of interpretable categories (topics).

Topic modeling is an unsupervised task, i.e. no manually annotated or labeled data is required as input. Topic models operationalise documents as *bags-of-words*, entirely detached from their within-document positions in the text and thus any representations of syntax. A generative process is assumed to have produced the texts, in which topics correspond to distributions of words in a document.

Blei describes the mechanism of LDA with an analogy, illustrated in Fig. 4.1. Imagine reading a collection of documents with a set of colored markers. You highlight the words associated with a theme with the corresponding colour. Processing the entire collection of texts in this manner provides results analogous to that of LDA. Topics are the collections of highlighted words grouped by colour, each representing a different topic.

For each document, the topics it contains are analogous to proportions of highlighted colours.

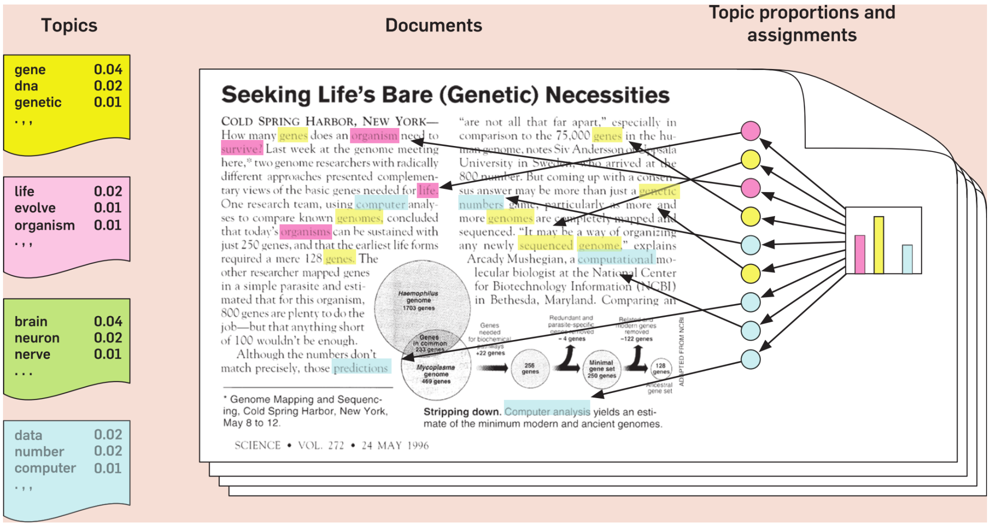


Figure 4.1. An intuitive explanation of topic modeling with LDA, reproduced from [22]

4.2 Topic modeling in practice

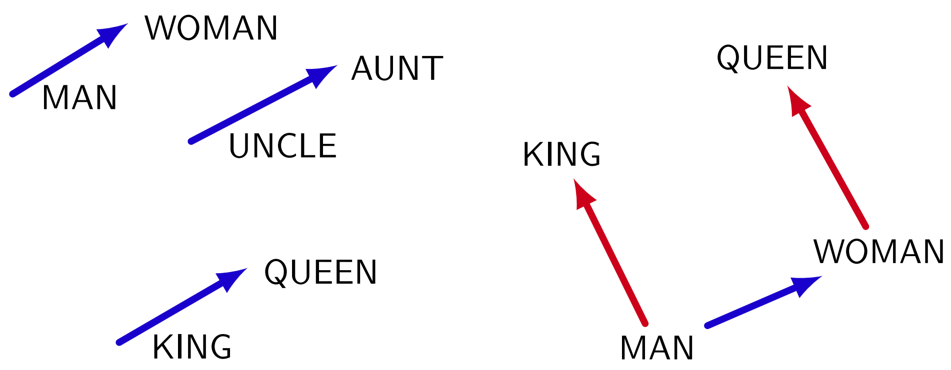
Nguyen, Liakata, DeDeo, *et al.* [4] describe a difference in approaches that often characterise interdisciplinary cooperation between the computer scientists familiar with NLP methods and the *domain experts* looking to apply the methods in real-world situations. Whereas 'computational analysis of text [...] is insight-driven', 'computational text analysis [...] tends to focus on automating tasks that humans perform inefficiently'. In the latter case, success is measured by quantitative performance, such as prediction accuracy in annotated datasets or in reaching state-of-the-art results on metrics purported to correspond to human intuitions.

The development of novel approaches to topic modeling is aimed at addressing shortcomings of LDA. These include the capability of modeling semantic aspects and being able to process short texts or small collections thereof.

4.3 Approaches to text representation

A prerequisite for all NLP tasks is the ability to represent words and documents as numeric vectors. The simplest approach is called One-hot encoding, in which each unique word in the vocabulary is assigned a column in the dataset with the value 1. A drawback of this is the impracticality of scaling to larger texts and vocabularies. The vectors for all words in the one-hot vocabulary are also orthogonal to each other, so there is no notion of some words being more semantically similar to each other than others.

An embedding is any method that maps a unit of text from the high-dimensional discrete matrix to lower-dimensional vector space. An example of this approach is word2vec [2], the algorithm discussed in the context of the game in Chapter 1. The computed word vectors retain semantic information and are compatible with basic vector operations, as illustrated in Fig. 4.2. Drawbacks include losing representations for different meanings of a word, the inability to handle OOV words as well as losing representations at the sub-word (token) level.



(a) An illustration of a 'gender dimension' represented in word vectors. **(b)** An illustration a the vector operation for analogical reasoning $X = \text{woman} + \text{king} - \text{man} \approx \text{queen}$

Figure 4.2. Simple vector operations can be used on Word2vec embeddings

4.4 Pre-trained language models and BERT

The innovation of language models marked the beginning of a new era in what NLP methods could accomplish. Given a sequence of words as input, they produce a completion, i.e. plausibly predict how the text might continue.

BERT is a transformer-based language representation model introduced by Devlin, Chang, Lee, *et al.* [23]. The core innovation behind transformers is that the entire sentence is processed as a whole, meaning that the word representations are informed not just by its preceding words but the overall context in the sentence. Modeling sentence structure as a whole is made possible by the self-attention mechanism, illustrated in Fig. 4.3. Note that the representation of the word *it* is influenced by the earlier word *animal*.

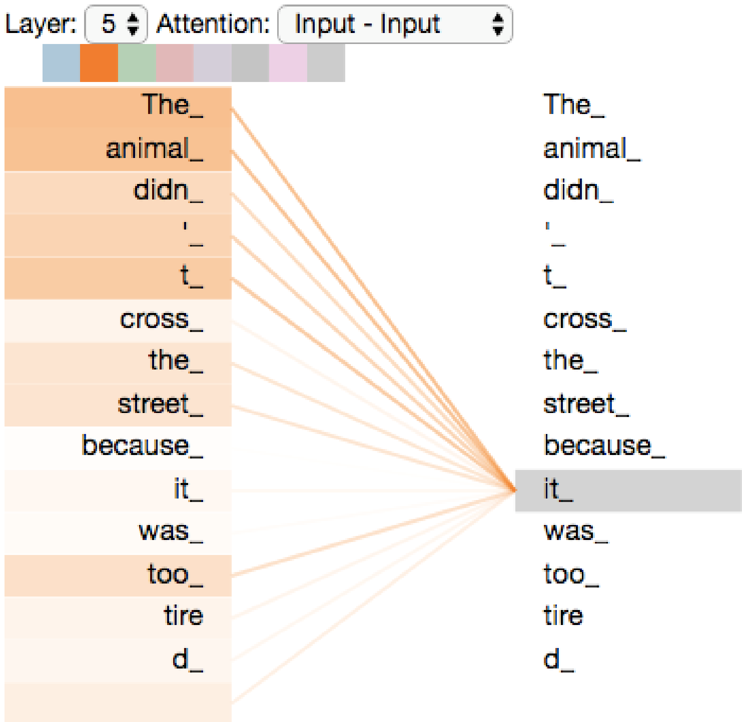


Figure 4.3. The self-attention mechanism of BERT. Figure from Jay Alammar [24]

The two main BERT training tasks are Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM is a 'fill in the blank' task, with the model predicting what a masked (hidden) word is likely to be. NSP is a binary classification task; given sentences A and B, predict whether B comes after A in the original text.

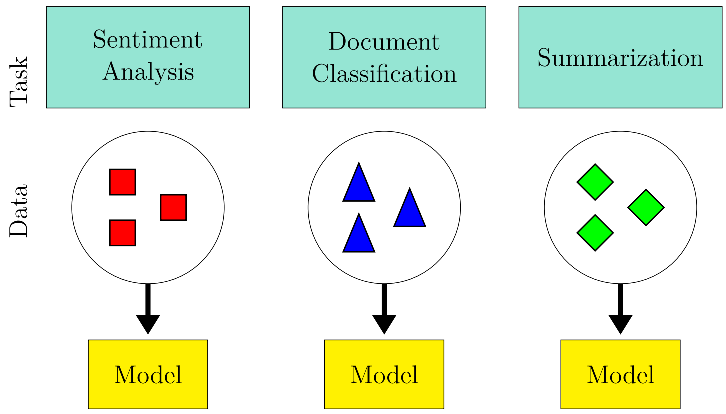
The input for BERT is a sequence of tokens. Tokens are pieces of words: if no match is found for an entire word in the model, the word is tokenized by subword segmentation of its constituent parts.

BERT processes the input through a sequence of multiple layers, starting with the pretrained Wordpiece embeddings. As the end result, the output of BERT is a sequence of vectors, one for each input token. Different layers of the BERT output have been shown to contain different aspects of linguistic information [25]. In the implementation used here, the model consists of 12 intermediate (hidden) layers.

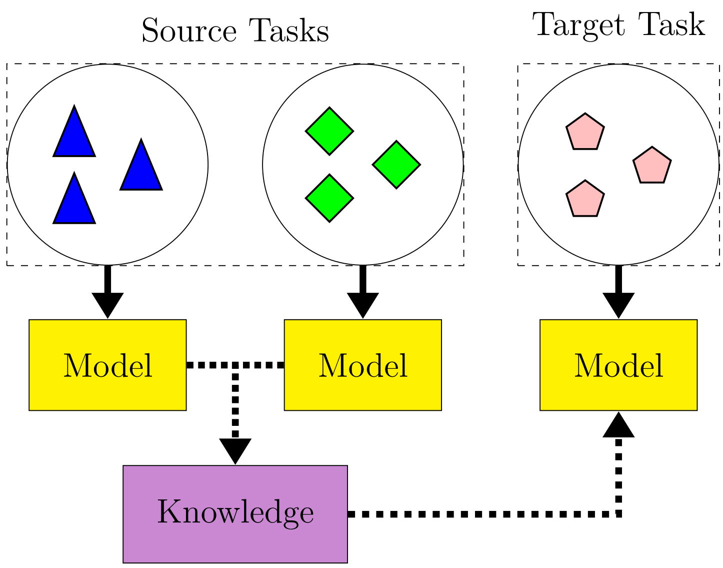
4.5 Transfer learning

Using BERT to produce completions to sentences or filling in blanks is the 'standard' use case. During training, the model learns to improve its predictions by comparing them to how the sentences continued in the original data (labels) (as illustrated in Fig. 4.4a).

In NLP transfer learning a language representation obtained after training for one task can be used as knowledge for a novel target task (see Fig. 4.4b). For instance, the vector representations of text that BERT produces to solve the predict-the-missing-word task could be repurposed for other tasks.



(a) Traditional machine learning in NLP



(b) Transfer learning in NLP

Figure 4.4. In transfer learning, representations learned for one task can be used as knowledge for another target task. Figures adapted from NLP Zurich [26]

4.6 Embedding clustering for topic modeling

The intersection of word embeddings and probabilistic topic modeling was first explored in a synergistic manner; embeddings were incorporated into the traditional topic modeling framework as supplementary data points. More recently, approaches have been proposed that entirely forego the probabilistic steps [27]–[29].

The similarity between probabilistic topic modeling and clustering was noted by Thompson and Mimno [28]. They argue that 'a clustering of tokens based on BERT vectors is functionally indistinguishable from a Gibbs sampling state for LDA' and extend the analogy to how these models represent local context and global information.

Sia, Dalmia, and Mielke [27] perform k-means clustering on word embedding vectors computed with various methods, including contextualised representations from BERT and the non-contextual word2vec. In their experiments on different variants, one or more of the versions outperformed LDA on evaluation metrics of topic coherence and diversity. However, none of the models performed consistently better. In a qualitative assessment, the BERT model was deemed as more 'syntactically-aware'.

The method for computing topic keywords in Sia, Dalmia, and Mielke [27] and Angelov [30]'s Top2Vec is based on cluster centroids and reranking words based on frequency. Grootendorst [29] challenges the assumption that the words near a cluster centroid will always be representative of a topic. For example, the word vector closest to the center point between 'cat', 'dog' and 'horse' is unlikely to be 'animal' or a similarly descriptive word.

Instead, Grootendorst [29]'s method treats the generation of topic representations for summarisation purposes as a separate process, leveraging a class-based version of Term Frequency-Inverse Document Frequency (TF-IDF). In addition to avoiding the 'centroid assumption', this separation of topic generation and representation implicitly enables the use of sentence-level representations in clustering.

4.7 Computational methodology and theory

The computational analysis forming the experimental part of this thesis is broadly similar to that of Grootendorst [29]'s BERTopic outlined above. The theoretical groundwork for computing the embeddings was described in Section 4.4. The method used for dimensionality reduction is discussed below, followed by the principles behind the clustering algorithm and finally with that of the cluster labeling process.

4.7.1 Dimensionality reduction with UMAP

Dimensionality reduction is a technique for constructing a lower-dimensional representation of a high-dimensional dataset. It can be used as part

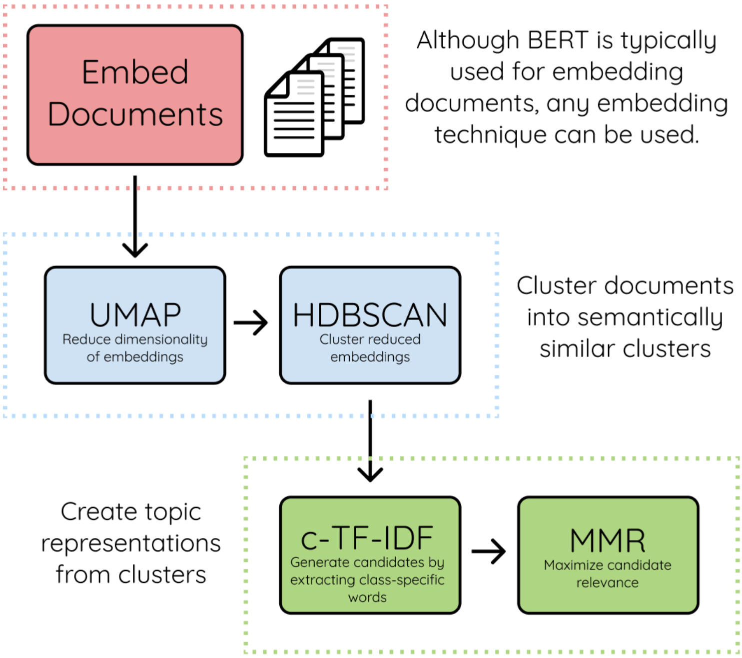


Figure 4.5. Topic modeling with BERTopic consists of three steps; computing document embeddings, reducing their dimensionality with UMAP, clustering with HDB-SCAN and creating topic descriptions in the form of keywords. Figure from Grootendorst [29]

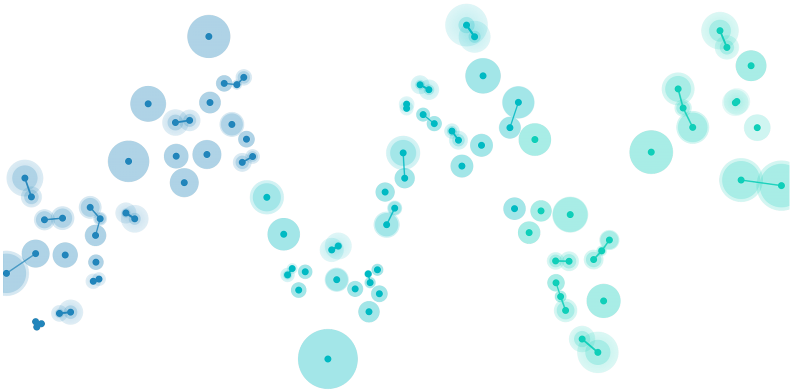
of the explorative step of analysis for visualisation and summarisation purposes or as an intermediate step as an input to another model. Dimensionality reduction algorithms can be broadly categorized into those based on matrix factorisation, such as Principal component analysis (PCA), and those involving graph layout methods. [31]

Uniform Manifold Approximation and Projection (UMAP) is a relatively novel graph layout based dimensionality reduction technique by McInnes, Healy, and Melville [32]. Compared to t-distributed stochastic neighbor embedding (t-SNE), one of the widely used state-of-the-art techniques, UMAP offers some key advantages such as faster computation especially on large datasets, better preservation of global structure and better usability. [33]

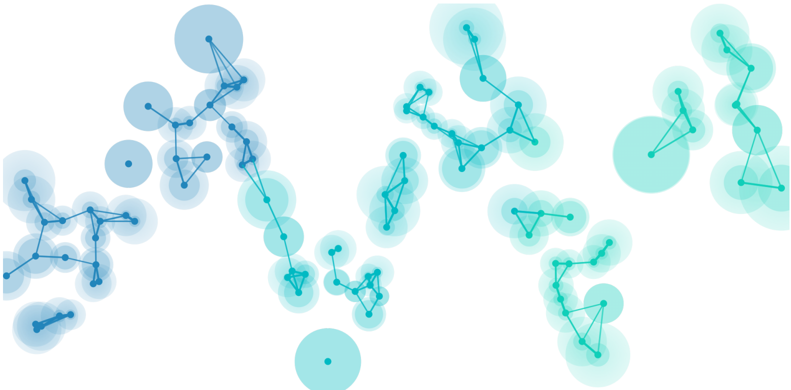
UMAP uses graph layout algorithms to construct a high-dimensional graph representation of the data, based on which a structurally similar lower-dimensional projection is constructed.

The first step is the conversion of the data into a weighted graph representation called a fuzzy simplicial complex, in which edge weights represent the likelihood of two points being connected.

Connectedness is defined as overlaps between radii centered on individual data points called 0-simplices. Determining the appropriate extent of the radius is computed locally based on the distance required to reach each point's nth nearest neighbor. The likelihood of connection between points



(a) Construction of the fuzzy simplicial complex with `n_neighbors = 2`



(b) Construction of the fuzzy simplicial complex with `n_neighbors = 5`

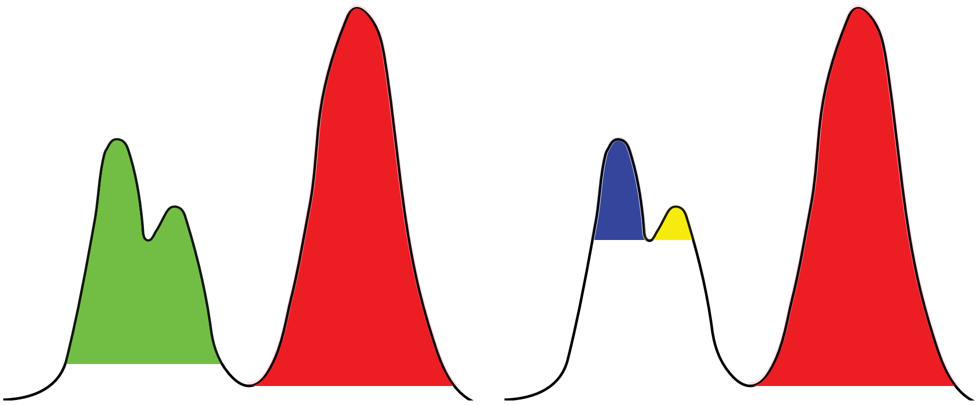
Figure 4.6. Fuzzy simplicial complexes in UMAP

decreases with increasing distance and the number of neighbors reached, becoming fuzzier as more neighbors are reached. Sets of points connected by extended radii are called n-simplices.

After the fuzzy simplicial complex representing the high-dimensional data is computed, it is projected into the desired lower-dimensional representation with a force-directed graph layout algorithm. This step makes use of stochastic gradient descent to minimize the difference between respective distances in the high-dimensional data and low-dimensional projection.

4.7.2 Clustering with HDBSCAN

Density-based clustering can be understood as 'finding islands (clusters) in a sea (of noise)'. From a probabilistic perspective, clusters are dense (highly probable) regions of data, separated by sparse (improbable) regions. The clusters can be assigned using the Probability density function (PDF) by setting a global threshold value λ and designating the distinct peaks above it as separate clusters (Fig. 4.7).



(a) Two distinct clusters, obtained with a lower λ

(b) Three distinct clusters, obtained with a higher λ

Figure 4.7. Clusters with different threshold values λ

To deal with peaks of different magnitude, ie. clusters of varying density, the threshold value is determined locally. In the example presented in Fig. 4.7, the overlapping area under the blue and yellow peaks (right) is greater than the two peaks' total area, so they are treated as a single green cluster (left) with bumps of noise instead of distinct clusters.

Formally, the area of the PDF regions is called persistence, determined by a method called excess of mass (eom). The objective in clustering is to maximize the total sum of cluster persistence, subject to the constraint that clusters do not overlap. In practice, the underlying PDF of the data is not known and must be empirically estimated.

The local density is estimated using core distance, defined as the distance to the kth nearest neighbor of a point. In other words, the core distance $\mathrm{core}_k(a)$ for point a is the smallest distance ε required to reach its k nearest neighboring points. As seen in Fig. 4.8, points in dense regions have small core distances and vice versa.

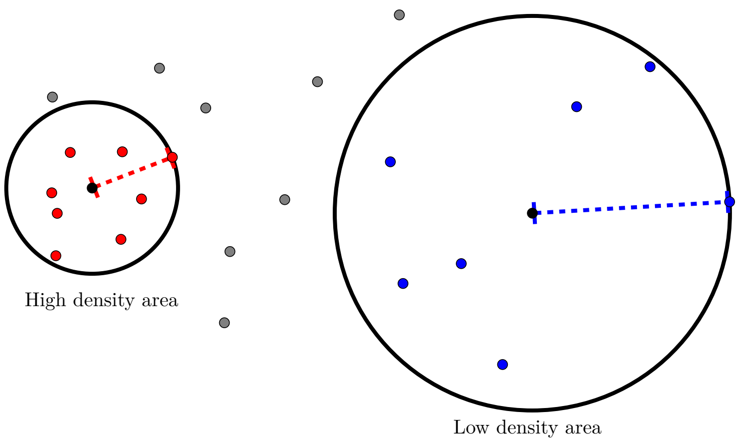


Figure 4.8. Core distances for $k = 7$ for example points in a dense region and a sparse region.

Distances between points are computed with a metric called the mutual reachability distance.

$$d_{\text{mreach-k}}\,(a,b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a,b)\} \tag{4.1}$$

where $d(a,b)$ is the pre-computed distance between points a and b and $\text{core}_k(a)$ and $\text{core}_k(b)$ are the core distances of points a and b. In effect, mutual reachability distance acts as a transformation for the data by repelling points in sparse regions from each other.

This means that for points to be regarded as connected, they must

- Be located in a dense enough region
- Be close enough to each other

After this transformation, consider the data as a weighted graph, with vertices corresponding to each point and edges between points a and b with weight $d_{\text{mreach-k}}\,(a,b)$. Based on this graph, the minimum spanning tree (MST) is computed by connecting all points with the minimum possible distance (Fig. 4.9).

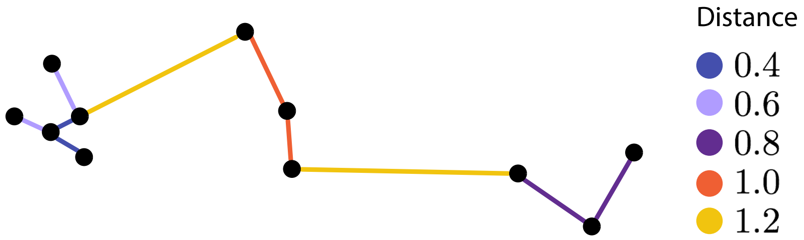


Figure 4.9. The minimum spanning tree.

The MST is then pruned by merging branches that contain fewer points than the desired minimum cluster size to their parent branches to obtain the single linkage tree (Fig. 4.10).

To determine the final clustering, the stability of each cluster in the tree is calculated with

$$\sum_{p \in \text{cluster}} (\lambda_p - \lambda_{\text{birth}}) \tag{4.2}$$

where λ_{birth} is the threshold λ beyond which the cluster splits and λ_p is the threshold beyond which point p no longer belongs to the cluster.

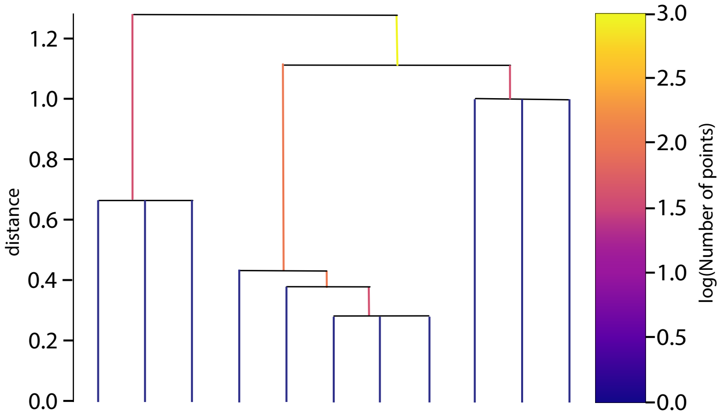
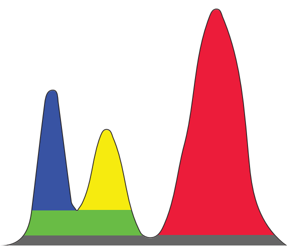
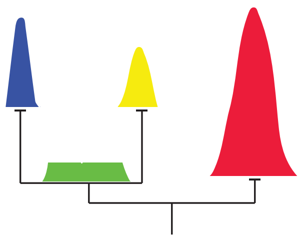


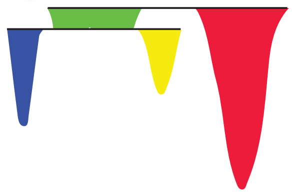
Figure 4.10. The single linkage tree.



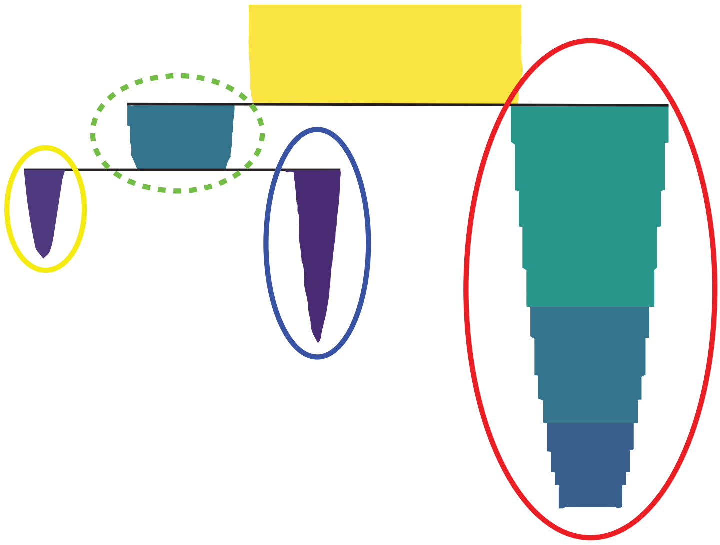
(a) The density plot . . .



(b) . . . can be visualised as a hierarchy . . .



(c) . . . and is upside down by convention.



(d) The condensed tree plot from HDBSCAN for the same simulated data, annotated with the blue, yellow and red clusters. The green area overlaps with its child clusters and is thus labeled noise.

4.7.3 Topic representation

Topic representation here refers to computing how a topic is presented to the user, as opposed to the topic generation referring to the output of the clustering algorithm. *Separation* of the two means that topic representation is only influenced by the sentence-topic assignments, whereas the embeddings or their lower-dimensional transformations are not considered.

The algorithm used is the class-based variant of TF-IDF proposed by Grootendorst [29]. Whereas the standard TF-IDF represents the importance of a word to a *document*, the variant considers the importance of a word to a *topic*. This change is quite simple to implement in practice; documents in each cluster are concatenated together, resulting in one 'pseudodocument' per cluster. The weight W, ie. relevance, of term t for class c is

$$W_{t,c} = \underbrace{\mathrm{tf}_{t,c}}_{\text{term frequency}} \cdot \underbrace{\log\left(1 + \frac{A}{\mathrm{tf}_t}\right)}_{\text{inverse class frequency}} \tag{4.3}$$

Term frequency $\mathrm{tf}_{t,c}$ is the proportion of how many times term t appears in class c out of its total terms. In the computation of inverse class frequency, A is the average number of terms per class and term frequency tf_t is the frequency of term t across all classes.

The topic representation methodology is extended from the one in [29] with the precomputed PoS and NER labels (Section 3.2.2). This extension of the method is implemented by constraining the terms to inlcude only those of the desired part-of-speech or named entity category before the concatenation of sentences by cluster.

5. Analysis

5.1 Formulation of research questions

RQ1. Does clustering sentence embeddings from BERT produce meaningfully interpretable categories?

   (a) Are the clusters discriminated by semantic, syntactic or other aspects?

RQ2. Based on the results of the topic modeling, which themes are present in the news coverage of Huawei?

   (a) How have Finnish institutions approached mitigating national security risks associated with a foreign vendor of 5G technology?

   (b) Are there connections to themes discussed in cybersecurity strategies in the news coverage?

5.2 Overview of the methodology

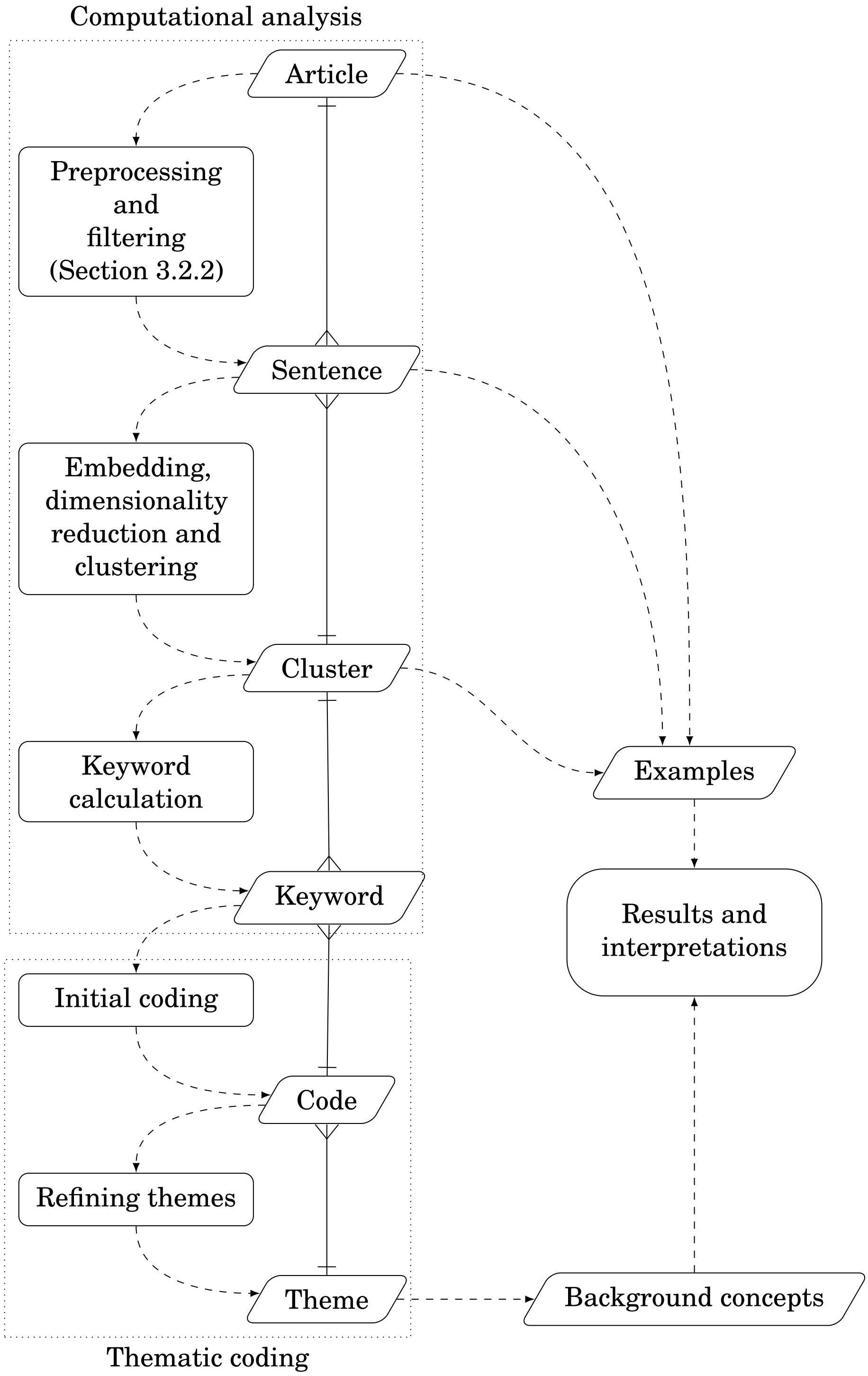


Figure 5.1. The left-hand side represents the main processing steps, both computational and manual, and the middle column represents data or other units of meaning being transformed.
An informal notion of cardinality indicates the relationships between data; an article contains multiple sentences, each of which belongs to one cluster. A single cluster has multiple keywords, which are grouped as codes and again into themes.

5.3 Computational analysis

The computational steps of the analysis were implemented in Jupyter notebooks using the Python programming language. The main steps were the computation of sentence embeddings, the dimensionality reduction and the clustering.

5.3.1 Sentence embeddings

Embeddings were computed using the BERT model by Virtanen, Kanerva, Ilo, *et al.* [34] [1]. The embeddings were processed using a method adapted from Chris McCormick [35]. Details on the processing of BERT outputs can be found in Appendix A.

5.3.2 Dimensionality reduction

Dimensionality reduction of the 768-dimensional text embedding vectors was performed using UMAP, introduced in Section 4.7.1. UMAP is applied twice for each collection of embedding vectors; one projection into 2-dimensional space for visualisation purposes. A higher-dimensional (specified using the parameter `n_components`) projection is used as the input for the clustering step.

The number of neighbors (`n_neighbors`) parameter controls the influence of local and global structure in the data. In practice, a low value will result in preserving the relationship between nearby data points and a large one in the 'big picture'. Other parameters held equal, a lower value will usually result in more topics and vice versa. Minimum distance (`min_dist`) is the smallest permissible distance between distinct data points.

Angelov [30] reports that using `n_neighbors` $= 15$ and `n_components` $= 5$ most consistently produced good results.

5.3.3 Clustering and keywords

Five-dimensional data from UMAP were clustered using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). HDBSCAN is a density-based clustering algorithm with properties that are favourable for the task at hand. It does not require specifying the number of clusters beforehand and is indifferent to cluster shape.

The most important parameter to influence clustering results is the

[1] Available from `https://huggingface.co/TurkuNLP/bert-base-finnish-cased-v1`

minimum cluster size (`min_cluster_size`). As suggested by the name, its value should be set to the smallest grouping of data points that should be considered a cluster.

HDBSCAN does not attempt to force every data point into a cluster but can instead designate them as outliers. This can be influenced by varying the minimum samples (`min_samples`) parameter value, which is by default set to equal `min_cluster_size`. Theorywise, this is related to the computation of the mutual reachability distance described in Section 4.7.2. Lower values will assign points that would otherwise be labeled noise to their most similar neighboring cluster instead.

The 10 most representative nouns, verbs and adjectives from each cluster and were used as cluster keywords, based on their TF-IDF weight scores.

5.4 Interactive exploration for parameter tuning

A graphical user interface was implemented with the Streamlit library [2] and used for interactive visualisation and exploration of the sentence embeddings. Example screenshots of the software can be found in Appendix C The different parameters available for user fine-tuning, their significance and the values used for obtaining the final results are presented in Table 5.1.

Parameter	Purpose	Value
text_search_string		huaw
embedding_layer	Hidden layer from BERT to use	12
max_sentence_chars		100
min_sentence_chars		0
n_neighbors	Balance between local and global structure	30
min_dist	Distance between similar points	0.0
densmap	Supplementary technique for preserving local structure	FALSE
n_components	Dimensionality of data used for clustering step	5
min_cluster_size	Smallest grouping desired to form a cluster	10
min_samples	Influences ‘conservativeness’ of clustering and threshold for assigning points as noise	5
cluster_selection_epsilon	Threshold for joining neighboring clusters together	0.3
cluster_selection_method		leaf
n_features	Consider the n most frequent terms	1000
ngram_range	Length of n-grams to include	(1, 2)
min_df	Ignore rarest terms	0.05
max_df	Ignore the most common terms	0.95
n_top_words	Number of keywords to output	10

Table 5.1. After experimentation with the interactive interface, this configuration of parameters was selected

[2] `https://docs.streamlit.io/`

5.5 Thematic analysis

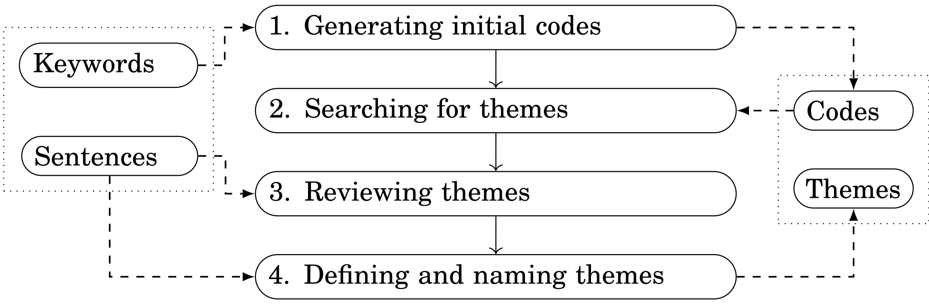


Figure 5.2. Structured process used in the thematic analysis of the clusters

Thematic analysis was used as the first step of qualitatively analysing the results. It has been described as a 'powerful and flexible method of qualitative analysis' and argued to be 'foundational for other qualitative research methods' [36].

In this context, *theme* is defined as a 'patterned response or meaning derived from the data that informs the research question'. The process of thematic analysis implemented here, as described by Kiger and Varpio [36], is illustrated in Fig. 5.2.

During the initial coding step, the objective is the generation of *codes* that act as preliminary scaffolding for further refinement. This step was informed through the definition of codes as '[t]he most basic segment, or element, of the raw data or information that can be assessed in a meaningful way regarding the phenomenon'. Contrasted with themes, codes are less abstract and have a tighter connection to the raw data. The keywords for each cluster were used as raw data.

After the initial coding, the search for themes emerges. The first pass is informed by the codes generated in the previous step, before expanding the analysis to the sentences themselves. Finally, the themes are named and defined, considering both the source sentences and keywords they emerged from as well as the themes' relationships to each other.

6. Results and interpretations

In this chapter, the results are presented. From the computational analysis, key descriptive statistics are given, followed by a description of the intermediate and final outputs of the structured thematic analysis. The results are examined in consideration of the theoretical background concepts.

6.1 Descriptive statistics of the results

The sentence length constraints resulted in a total of 508 sentences from 175 different articles.

Clustering using the parameters choices described in Table 5.1 produced 13 clusters containing 346 (68.1 % of total) sentences, with the remaining 162 (31.9 % of total) classified as noise.

The count of sentences in each cluster is presented in Fig. 6.1. Two clusters had 10 sentences, the minimum allowed with the `min_cluster_size` parameter, with the median being 29 and a maximum of 46.

After exploring the results, the inclusion of named entities as cluster keywords was deemed ineffectual, as they were too infrequent across most clusters. Instead, a summary of their frequency across the whole of the articles examined is presented in Table 6.1.

6.2 Thematic coding of the cluster keywords

The first step of the thematic analysis was processing cluster keywords into codes. The keyword lists and the full audit trail describing the code assignments and their justifications for each cluster can be found in Appendix D.1.

The initial coding of the keywords resulted in a total of 78 keywords (73 unique) grouped into 27 codes. For two clusters, no sensible coding was

Type of named entity	Number of entities
ORG	3033
GPE	1335
NORP	473
PERSON	439
DATE	386
PRODUCT	231
LOC	124
PERCENT	65
MONEY	30
TIME	12
QUANTITY	8
EVENT	8
WORK_OF_ART	7
FAC	7
LAW	5
LANGUAGE	2

Table 6.1. Counts of named entity types in the articles

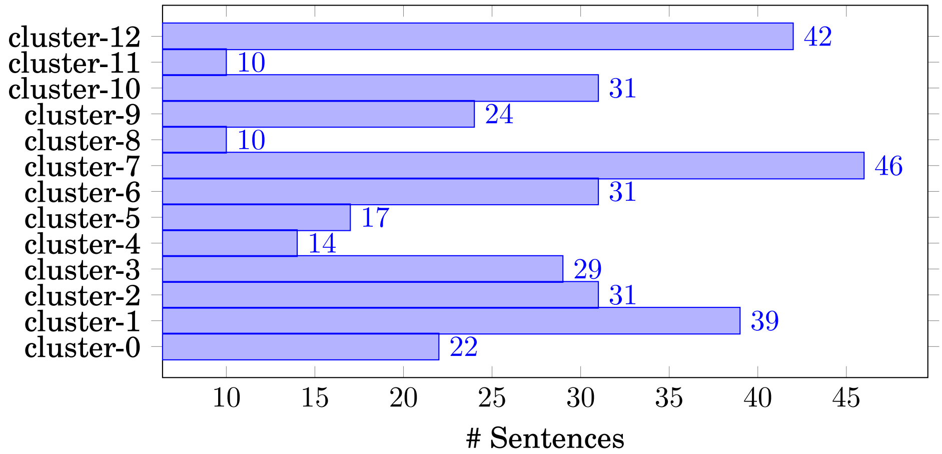


Figure 6.1. Number of sentences in each cluster

deemed possible due to the keywords being too general to separate from the other clusters.

The grounds for code assignment varied between clusters. In the simplest cases, the coding was based on grouping synonymous or related words for physical objects.

6.3 Refining codes to themes

The 27 codes were categorized into three top-level themes: *Accusations of misconduct*, *Devices and Products* and *The Telecommunications Market*. In the following sections, these themes are described and interpreted.

6.3.1 Accusations of misconduct

Cluster	Code	Word
1	cessation, termination, ending	ties
		to disappear
		to terminate
		to sever
	distress or trouble	to limp
		hard-pressed
4	assessing an evolving situation	to assess
		to monitor
		to keep an eye on
	communications and public messaging	report
		word
		communications
	expressing understanding	optimistic
		understandable
		cautious
	voicing concerns	accusation
		worry
6	expressing opinions	to comment
	vehement denial	to deny
		to rebut
		to reject
	voicing concerns	accusation
		claim
		suspiction
11	legal proceedings	court of justice
		lawsuit
		not guilty
		to prosecute
		to sue
12	being in opposition to	opposed to
	expressing accusations	to accuse
	expressing concern	worry
	expressing suspicion	to claim
		to suspect
	geopolitical significance	country
		the West
	imploring action	to pressure
		to warn

Table 6.2. Accusations of misconduct

The theme *Accusations of misconduct* contained discussion of various accusations directed at Huawei, the company's responses to them as well as commentary by third parties.

Contents of the allegations

The allegations cited include unspecified 'concern for national security', violating sactions against Iran [1], claims of the company collaborating with 'China', '[its] government', '[its] military' or '[its] security services' or even effectively acting as a state-owned company. [2, 3]

Information is mostly attributed to nonspecific United States sources, with only a couple mentioning a specific organizational entity, the National Security Agency. This type of discourse presents geopolitical security issues as a narrative, complete with personified actors, as described by Jansson and Sihvonen. Most of the sentences contain qualifiers to imply some degree of uncertainty. [4, 5]

The codes 'warning and pressuring others to take action' and 'geopolitical significance' arise from discussions of actions by the US taken to influence policy in allied countries. [6, 7]

Responses from the company

Codes from cluster 6 included 'vehement denial' from keywords 'to deny, to rebut, to reject' and the more neutral 'expressing opinions'.

Legal remedies

In early 2011, Huawei sued Motorola in a case related potential violation of trade secrets [8], and was successful in the case [9].

In 2019, Huawei sued the United States for enacting a law banning the federal government from cooperating with companies that cooperate with Huawei. The article notes that 'the option of seeking justice through an independent judiciary' is 'a means central to the world[-view]' in the Western world, but 'would be near impossible [in China]' [10]. This observation connects to the theme of legislation in the cyberstrategy classification, with the further implication that tensions can arise when an entity operates under multiple legal systems.

[1] https://yle.fi/uutiset/3-10544998
[2] https://yle.fi/uutiset/3-10791332
[3] https://yle.fi/uutiset/3-11640308
[4] https://yle.fi/uutiset/3-10591742
[5] https://yle.fi/uutiset/3-10703110
[6] https://yle.fi/uutiset/3-11417355
[7] https://yle.fi/uutiset/3-11021873
[8] https://yle.fi/uutiset/3-5311690
[9] https://yle.fi/uutiset/3-5342195
[10] https://yle.fi/uutiset/3-10677731

Contextualising commentary

Codes arising from cluster 4 keywords included the rather neutral 'cautiously assessing an evolving situation' and the more positive 'expressing understanding'. Opinions are cited from Finnish officials and academic experts, who are interviewed to explain the reasons behind strong rhetoric and restrictive actions by Western countries on Huawei. For example, an article from 2019 [11] was published as news broke that Sweden is debating legislation to give security authorities a veto on providers of 5G network technology.

A director from the cybersecurity centre of the Finnish Ministry of Transport and Communications explains that Finland can 'deter threats already within the current legislation', while conceding that banning an individual company is not possible. The rationale for judging existing powers as sufficient can be seen as technology-centric, pointing out that infrastructure providers are 'obliged to report any identified shortcomings in information security' and 'comply with regulations'. The official also expresses a wish to establish a 'pan-European solution'. This is because 'many novel and even societally critical services' will rely on 5G, so 'ensuring [their security] with all available means is of utmost importance.'

Another ministry official declines to identify any individual company or country by name, as '[it would be] contrary to our tradition of the rule-of-law'. A professor of network technology interviewed comments on the issue being 'partly related to superpower politics'.

[11] `https://yle.fi/uutiset/3-11018309`

6.3.2 Devices and products

Cluster	Code	Word
2	devices and technical matters	phone
		technical
		upgrade, update
3	assessing quality	expensive
		to test
		unnecessary, useless
	mobile phones	display (of a device)
		mobile phone
		phone
		service provider
		to release

Table 6.3. Devices and products

In this theme, the articles mentioning Huawei predominantly cater to consumer interests, covering device reviews and product launches. The intersection of geopolitical issues and the consumer viewpoint manifests as articles clarifying the potential security concerns the practical impact of sanctions on functionality. Using the open source Android operating system in itself cannot be restricted [12, 13]. Due to restrictions in accessing the Google Play store, the contact tracing app Koronavilkku is not available on newer Huawei models. [14]

Controversies have had detrimental effects on Huawei's business. [15] Uncertainty erodes trust, even in the absence of concrete evidence. [16] More concretely, restrictions on Huawei's use of the Android operating system have an impact on user experience. [17] Competitors, including Nokia, are assessed to be potential benefactors. [18, 19]

[12] https://yle.fi/uutiset/3-10791332
[13] https://yle.fi/uutiset/3-11525049
[14] https://yle.fi/uutiset/3-11525049
[15] https://yle.fi/uutiset/3-8484078
[16] https://yle.fi/uutiset/3-10945853
[17] https://yle.fi/uutiset/3-10794830
[18] https://yle.fi/uutiset/3-10667311
[19] https://yle.fi/uutiset/3-11619907

6.3.3 The Telecommunications Market

Cluster	Code	Word
5	emphasising quality	an investment
		good
		good-quality
		quality
		strong
		to invest in
	markets and their interconnectedness	ecosystem
		the market
7	markets and competition	competitor
		manufacturer
		profit
		the market
		to compete
	taking over in a competitive situation	neck
		to breathe
		to gasp
	success in competition	number one
		the top
		to grow
		to rise
		to succeed
9	economic benefits	to invest in
	research and development	development
		research
	the job market	to employ
		to hire

Table 6.4. The Telecommunications Market

Huawei's rise to dominance

The rapid growth of Huawei's market share in both network infrastructure and consumer smartphones is emphasised, for example noting that 'Huawei is an exceptional success story' [20] and noting their 'market dominance' [21].

References are made to both Western and Chinese companies, including Swedish Ericsson, French Alcatel-Lucent, American Apple and Chinese Huawei, ZTE, Lenovo and Xiaomi. The majority of references to Chinese companies contain a mention to their nationality (e.g. 'the *Chinese* company Huawei').

[20] https://yle.fi/uutiset/3-10677886
[21] https://yle.fi/uutiset/3-10667311

The Finnish company Nokia is often mentioned in comparison to Huawei and other players, often in an unfavourable terms. The headline of an article from 2019 [22] posits that 'Nokia deserved a spanking' after a dramatic decline in stock value. The decisions to merge with the Siemens network division and to acquire Alcatel-Lucent are presented as poor investments made at the expense of neglecting research and development. In contrast, 'Huawei's yearly spending on development is over three-fold compared to Nokia' and boasts having 'more 5G patents than anyone else'.

In addition to R&D efforts, the vast size of the Chinese domestic market is credited as a catalyst for growth. [23]

Economic benefits of investments

Huawei's research and development efforts are discussed, for example the 'research and product development departments in Helsinki and Tampere' and the economic and job market benefits they bring [24]. Mentions of other companies with research units in Finland include Samsung, LG and Intel.

[22] https://yle.fi/uutiset/3-11037795
[23] https://yle.fi/uutiset/3-11018309
[24] https://yle.fi/uutiset/3-9654499

7. Discussion

In this chapter, the methodology is assessed, including the qualitative thematic analysis framework and the features of the clustering results. Contrasting attitudes to the nature of topic modeling are discussed, as well as recent criticisms on the nature of capabilities of pretrained language models.

7.1 Expanding the modalities for semantic interaction

In the opening chapter, the game Semantle was used as an example of how the relatively antiquated word2vec algorithm can provide a satisfying user experience with a cleverly designed interface. Since the model only reported similarity scores for words after the player had input them, the words and scores actually visible to the user represented only the curated tip of the iceberg of the vocabulary.

The framework developed for this thesis was also developed with the aim of being tolerant to individual, or even prevalent failure cases of embeddings carrying no information meaningful to the user. This was addressed in two stages. At the computational level, sentences assessed as noise by Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) were excluded before rerunning the remaining data. Similarly, the option to manually exclude unwanted clusters was available to the user.

Further work on improving the human-machine interaction is warranted, including novel ways to incorporate user feedback to steer the model to suit the needs of the user. An interesting example of this paradigm known as semantic interaction is described in work by Gonzalez Martinez [37]. The method developed aims to make the link between the high-dimensional embedding vectors and their reduced 2D visualisations bidirectional by training a deep surrogate network to approximate their mapping in a linear, and thus reversible, fashion. In practise, this allows a user to direct the model in the visual space by moving points in a drag-and-drop manner, with the model incorporating these adjustments.

7.2 Interpreting the clusters

A hands-on application of using sentence embedding for topic modeling was informative in demonstrating that, at least in this case, the output contained a mix of syntactic and semantic relations. Special attention had to be given in the qualitative part of the analysis to enable a coherent interpretation of the findings, as the raw cluster outputs were not suitable for relating to the background theoretical concepts.

Several ways of indicating a direct quotation or transmitting the thoughts of an interviewee were present in cluster 4. Some sentences feature a quotation dash to indicate this [25] [26], whereas paraphrased thoughts are indicated by verbs such as *say*, *stress refer*, *assess* or *tell*.

Demonstrating the potential of the method to address contextual meaning can be seen in the issue of metaphorical phrases. In cluster 7 containing discussion of the market situation, three sentences described the idea of taking over in competition with the expression 'breathing/panting down [someone's] neck'.

The degree to which clusters could be subjectively placed on a 'syntactic-semantic spectrum' also varied. The clearest case of a cluster being best discriminated from others by syntax was seen in the case of questions. A simple string match checking for the presence of a question mark in the sentence verified this. Cluster 0 contained 19 out of the total 23 such sentences in the clustered dataset. The remaining 4 were in cluster 1, perhaps due to it being closest to cluster 0 in the hierarchical tree.

Verifying whether the presence of a symbol such as a question mark is connected to the clustering of sentences is one of the simplest approaches to assessing hypotheses and observations that arise when interacting with the model. For more nuanced cases, for instance assessing evaluating whether quotations or metaphors do indeed tend to be discriminated by the clustering algorithm, more elaborate investigations are needed. Formal evaluation could make use of 'probing datasets', ie. text collections that

[25] `https://yle.fi/uutiset/3-10546742`
[26] `https://yle.fi/uutiset/3-10546742`

have been collected or constructed to contain examples and counterexamples of the linguistic feature of interest.

7.3 What is a topic?

In a comprehensive review on the evolution of topic modeling methods, Churchill and Singh [38] are skeptical of whether topic modeling with pretrained embeddings is feasible without including a probabilistic, LDA-like step. The central concern is the difference in objective, with the former being '[an] attempt to cluster words based on their contextual similarities' instead of '[producing] a set of words related to a topic'. Furthermore, 'language models like BERT are trained to produce coherent sentences and syntax, which is not required in topics'.

Taking into account the objective function used for training the model is an important aspect and warrants further study, including the development of Pretrained language model (PLM)'s trained with a semantic similarity objective. However, implying that the principles of traditional probabilistic topic are wholly aligned with subjective assessments and practical use cases is an open question.

From a humanities scholar's perspective, as sardonically noted by Klein, Eisenstein, and Sun [39], 'there is no inherent reason to believe that words grouped together on the basis of co-occurrence statistics should really mean or prove anything, aside from the winkingly suggestive similarities that these word groups so often display.'

Pääkkönen and Ylikoski [40] describes a dichotomy of approaches to topic modeling; topic realism and topic instrumentalism. In topic realism, the task involves classification of the data into fixed categories. Topic instrumentalists, on the other hand, regards the modeling task as clustering, with the aim of arranging the data into groups that contain similar items on any dimension.

In this thesis, an approach leaning towards topic instrumentalism is more applicable. The goal is, in fact, identical to that expressed by Blei in the context of the initial LDA topic model: '[a] browsing experience' and 'an algorithmic solution to managing, organizing, and annotating large archives of texts' [22].

	Topic realism	Topic instrumentalism
Topics	correspond to underlying theoretical concepts	are clusters of words that reappear across texts
Concepts	actually 'exist' in the text	depend on methodological and theoretical choices
Topic modeling	objectively measures the concepts	can guide further interpretation

Figure 7.1. A comparison of two approaches to topic models, proposed by Pääkkönen and Ylikoski [40]

7.4 Can language models cheat?

Deep learning based techniques excel at finding such patterns and regularities in training data that are able to make inferences on unseen test data. Exactly what the model bases its representations and predictions on requires additional investigation. One commonly employed method involves the crafting of *probes*, deliberately crafted evaluation tasks that aim to diagnose the model's capabilities.

McCoy, Pavlick, and Linzen [41] devised a probing task called HANS designed to impede the use of certain 'syntactic heuristics'. They are superficial shortcuts that can nevertheless perform well on commonly used benchmarking tasks. After evaluating a BERT model on the dataset tweaked to render the heuristics useless, its performance dropped dramatically. Work by Niven and Kao [42] provides another example. After first reaching performance almost equal to the human baseline on an argument comprehension benchmark, the dataset was then tweaked to preclude certain 'spurious statistical cues' from informing the result. Again, it turned out that this change dropped the performance to near chance accuracy.

7.5 Concluding remarks

In the opening chapter of this thesis, I hoped to convey some basic intuition about what word2vec embeddings are with a few paragraphs of lightweight remarks on a simple online game, before advancing into the technical details and process of clustering sentence embeddings from BERT. Broadly speaking, the same methodology could be applied to embeddings from other language models, including perhaps the most well-known and much-hyped one known as Generative Pre-trained Transformer 3 (GPT-3).

The model, described by the philosopher David Chalmers as 'one of the most interesting and important AI systems ever produced' [43], is designed to provide a 'general-purpose "text in, text out" interface, allowing users to try it on virtually any English language task' [44].

After having experimented with the ideas that would lead me to pursue the ideas in this thesis for some time, I was delighted when OpenAI, the organisation behind GPT-3, announced in their blog in January 2022 the rollout of a new feature 'that makes it easy to perform natural language [...] tasks like [...] clustering [and] topic modeling' [45]. The announcement linked to an accompanying preprint [46], which I hoped might discuss the theoretical link between topic models and text embedding clustering. The only reference to clustering, however, was a matter-of-fact statement noting that '[t]hese embeddings are useful features for classification tasks and can also enable data visualisation applications via techniques such as clustering', with no mention of topic models. Evidently, I was not the only one making the connection between topic modelling and embedding clustering. Nevertheless it seems that the theoretical underpinnings of this remain to be clarified and explored in detail.

Bibliography

[1] T. Rizzo, 'Semantle is like the Dark Souls of Wordle', *Washington Post*, 4 April 2022, ISSN: 0190-8286. [Online]. Available: `https://www.washingtonpost.com/video-games/2022/04/04/semantle-wordle-hard-mode/` (visited on 10/06/2022).

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, 'Distributed representations of words and phrases and their compositionality', in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119. [Online]. Available: `https://arxiv.org/abs/1310.4546`.

[3] G. Smith. 'Semantle is hard mode Wordle, powered by a Google neural network', Rock Paper Shotgun. (8 February 2022), [Online]. Available: `https://www.rockpapershotgun.com/semantle-is-hard-mode-wordle-powered-by-a-google-neural-network` (visited on 30/03/2022).

[4] D. Nguyen, M. Liakata, S. DeDeo *et al.*, 'How We Do Things With Words: Analyzing Text as Social and Cultural Data', *Frontiers in Artificial Intelligence*, vol. 3, p. 62, 25 August 2020, ISSN: 2624-8212. DOI: `10.3389/frai.2020.00062`. [Online]. Available: `https://www.frontiersin.org/article/10.3389/frai.2020.00062/full` (visited on 08/09/2021).

[5] European Union Agency for Cybersecurity (ENISA). 'National Cybersecurity Strategies — ENISA'. (), [Online]. Available: `https://www.enisa.europa.eu/topics/national-cyber-security-strategies` (visited on 25/04/2022).

[6] J. Limnéll, K. Majewski and M. Salminen, *Kyberturvallisuus*. Jyväskylä: Docendo, 2014, 246 pp., ISBN: 978-952-291-047-9.

[7] K. Friis and O. Lysne, 'Huawei, 5G and Security: Technological Limitations and Political Responses', *Development and Change*, vol. 52, no. 5, pp. 1174–1195, Sep. 2021, ISSN: 0012-155X, 1467-7660. DOI: `10.1111/dech.12680`. [Online]. Available: `https://onlinelibrary.wiley.com/doi/10.1111/dech.12680` (visited on 01/04/2022).

[8] Huawei Cyber Security Evaluation Centre (HCSEC) Oversight Board, 'HCSEC Oversight Board Report 2018', Annual Report, Jul. 2018. [Online]. Available: https://web.archive.org/web/*/https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/727415/20180717_HCSEC_Oversight_Board_Report_2018_-_FINAL.pdf (visited on 04/11/2022).

[9] O. Lysne, *The Huawei and Snowden Questions*. Cham: Springer International Publishing, 2018, ISBN: 978-3-319-74949-5. DOI: 10.1007/978-3-319-74950-1. [Online]. Available: http://link.springer.com/10.1007/978-3-319-74950-1 (visited on 11/04/2022).

[10] K. Kaska, H. Beckvard and T. Minárik, 'Huawei, 5G and China as a Security Threat', *The NATO Cooperative Cyber Defence Centre of Excellence*, p. 26, 2019. [Online]. Available: https://ccdcoe.org/library/publications/huawei-5g-and-china-as-a-security-threat/.

[11] US House of Representatives, 'Investigative Report on the U.S. National Security Issues Posed by Chinese Telecommunications Companies Huawei and ZTE', 8 October 2012. [Online]. Available: https://intelligence.house.gov/news/documentsingle.aspx?DocumentID=96 (visited on 25/04/2022).

[12] DCMS, 'Huawei to be removed from UK 5G networks by 2027', Department for Digital, Culture, Media and Sport, Press Release, 17 Jul. 2020. [Online]. Available: https://www.gov.uk/government/news/huawei-to-be-removed-from-uk-5g-networks-by-2027 (visited on 25/04/2022).

[13] S. Jansson and T. Sihvonen, 'Kyberturvallisuus valtiollisena toimintaympäristönä ja siihen kohdistuvat uhkat', *Media & viestintä*, vol. 41, no. 1, 22 March 2018, ISSN: 2342-477X. DOI: 10.23983/mv.69950. [Online]. Available: https://journal.fi/mediaviestinta/article/view/69950 (visited on 06/10/2021).

[14] Yleisradio, 'Yle finnish news archive 2011-2018, source', Kielipankki, text corpus. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2017070501.

[15] ——, 'Yle finnish news archive 2019-2020, source', Kielipankki, text corpus. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2021050401.

[16] J. Kanerva, F. Ginter, N. Miekka, A. Leino and T. Salakoski, 'Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task', in *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium: Association for Computational Linguistics, 2018. [Online]. Available: http://turkunlp.org/Turku-neural-parser-pipeline/.

[17] K. Haverinen, J. Nyblom, T. Viljanen *et al.*, 'Building the essential resources for Finnish: The Turku Dependency Treebank', *Language Resources and Evaluation*, vol. 48, no. 3, pp. 493–531, 2014, ISSN: 1574-020X. DOI: 10.1007/s10579-013-9244-1. [Online]. Available: http://dx.doi.org/10.1007/s10579-013-9244-1.

[18] S. Pyysalo, J. Kanerva, A. Missilä, V. Laippala and F. Ginter, 'Universal Dependencies for Finnish', in *Proceedings of NoDaLiDa 2015*, NEALT, 2015, pp. 163–172. [Online]. Available: https://aclweb.org/anthology/W/W15/W15-1821.pdf.

[19] D. Zeman, M. Potthast, E. Duthoo *et al.*, *CoNLL 2018 shared task system outputs*, 2018. [Online]. Available: http://hdl.handle.net/11234/1-2885.

[20] J. Luoma, M. Oinonen, M. Pyykönen, V. Laippala and S. Pyysalo, 'A broad-coverage corpus for Finnish named entity recognition', in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, May 2020, pp. 4615–4624, ISBN: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.567.

[21] D. M. Blei and Andrew Y. Ng, 'Latent Dirichlet Allocation', *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2002. [Online]. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[22] D. M. Blei, 'Probabilistic topic models', *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, April 2012, ISSN: 0001-0782, 1557-7317. DOI: 10.1145/2133806.2133826. [Online]. Available: https://dl.acm.org/doi/10.1145/2133806.2133826 (visited on 05/04/2022).

[23] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', 24 May 2019. arXiv: 1810.04805 [cs]. [Online]. Available: http://arxiv.org/abs/1810.04805 (visited on 10/06/2021).

[24] Jay Alammar. 'The Illustrated Transformer'. (27 Jun. 2018), [Online]. Available: http://jalammar.github.io/illustrated-transformer/ (visited on 12/04/2022).

[25] A. Coenen, E. Reif, A. Yuan *et al.*, 'Visualizing and Measuring the Geometry of BERT', 28 October 2019. arXiv: 1906.02715 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1906.02715 (visited on 09/12/2021).

[26] NLP Zurich, director, *Thomas Wolf (HuggingFace): An Introduction to Transfer Learning and HuggingFace*, 17 March 2020. [Online]. Available: https://www.youtube.com/watch?v=rEGB7-FlPRs (visited on 02/11/2022).

[27] S. Sia, A. Dalmia and S. J. Mielke, 'Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, November 2020, pp. 1728–1736. DOI: 10.18653/v1/2020.emnlp-main.135. [Online]. Available: https://aclanthology.org/2020.emnlp-main.135 (visited on 02/09/2021).

[28] L. Thompson and D. Mimno, 'Topic Modeling with Contextualized Word Representation Clusters', 23 October 2020. arXiv: 2010.12626 [cs]. [Online]. Available: http://arxiv.org/abs/2010.12626 (visited on 06/09/2021).

[29] M. Grootendorst, *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*, 2022. DOI: 10.48550/ARXIV.2203.05794. [Online]. Available: https://arxiv.org/abs/2203.05794.

[30] D. Angelov, 'Top2Vec: Distributed Representations of Topics', 19 August 2020. arXiv: 2008.09470 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2008.09470 (visited on 20/09/2021).

[31] A. Coenen and A. Pearce. 'A deeper dive into UMAP theory'. (), [Online]. Available: https://pair-code.github.io/understanding-umap/ (visited on 07/03/2022).

[32] L. McInnes, J. Healy and J. Melville, 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction', 17 Sep. 2020. arXiv: 1802.03426 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1802.03426 (visited on 07/03/2022).

[33] A. Coenen and A. Pearce. 'Understanding UMAP'. (), [Online]. Available: https://pair-code.github.io/understanding-umap/ (visited on 07/03/2022).

[34] A. Virtanen, J. Kanerva, R. Ilo *et al.*, 'Multilingual is not enough: BERT for Finnish', 15 December 2019. arXiv: 1912.07076 [cs]. [Online]. Available: http://arxiv.org/abs/1912.07076 (visited on 03/01/2022).

[35] Chris McCormick. 'BERT Word Embeddings Tutorial · Chris McCormick'. (14 May 2019), [Online]. Available: https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/ (visited on 07/02/2022).

[36] M. E. Kiger and L. Varpio, 'Thematic analysis of qualitative data: AMEE Guide No. 131', *Medical Teacher*, vol. 42, no. 8, pp. 846–854, 2 August 2020, ISSN: 0142-159X, 1466-187X. DOI: 10.1080/0142159X.2020.1755030. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/0142159X.2020.1755030 (visited on 05/04/2022).

[37] A. Gonzalez Martinez, 'Exploratory Analysis of Research Publications with Human Steerable Black-Box Models. Towards Generalizing Inverse Computations for Semantic Interaction.', Thesis, University of Hawai'i at Manoa, 2021. [Online]. Available: http://scholarspace.manoa.hawaii.edu/handle/10125/75925 (visited on 19/09/2021).

[38] R. Churchill and L. Singh, 'The Evolution of Topic Modeling', *ACM Computing Surveys*, p. 3 507 900, 12 January 2022, ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3507900. [Online]. Available: https://dl.acm.org/doi/10.1145/3507900 (visited on 18/01/2022).

[39] L. F. Klein, J. Eisenstein and I. Sun, 'Exploratory Thematic Analysis for Digitized Archival Collections', *Digital Scholarship in the Humanities*, vol. 30, pp. i130–i141, suppl 1 1 December 2015, ISSN: 2055-7671, 2055-768X. DOI: 10.1093/llc/fqv052. [Online]. Available: https://academic.oup.com/dsh/article-lookup/doi/10.1093/llc/fqv052 (visited on 03/04/2022).

[40] J. Pääkkönen and P. Ylikoski, 'Humanistic interpretation and machine learning', *Synthese*, vol. 199, no. 1-2, pp. 1461–1497, December 2021, ISSN: 0039-7857, 1573-0964. DOI: 10.1007/s11229-020-02806-w. [Online]. Available: https://link.springer.com/10.1007/s11229-020-02806-w (visited on 03/04/2022).

[41] T. McCoy, E. Pavlick and T. Linzen, 'Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3428–3448. DOI: 10.18653/v1/P19-1334. [Online]. Available: https://aclanthology.org/P19-1334 (visited on 21/04/2022).

[42] T. Niven and H.-Y. Kao, 'Probing Neural Network Comprehension of Natural Language Arguments', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4658–4664. DOI: 10.18653/v1/P19-1459. [Online]. Available: https://aclanthology.org/P19-1459 (visited on 20/04/2022).

[43] D. Chalmers, 'GPT-3 and General Intelligence', *Daily Nous*, 30 Jul. 2020. [Online]. Available: https://dailynous.com/2020/07/30/philosophers-gpt-3/ (visited on 10/06/2022).

[44] OpenAI, G. Brockman, M. Murati and P. Welinder. 'OpenAI API', OpenAI Blog. (11 Jun. 2020), [Online]. Available: https://openai.com/blog/openai-api/ (visited on 10/06/2022).

[45] A. Neelakantan, L. Weng, B. Power and J. Jang. 'Introducing Text and Code Embeddings in the OpenAI API', OpenAI. (25 January 2022), [Online]. Available: https://openai.com/blog/introducing-text-and-code-embeddings/ (visited on 10/06/2022).

[46] A. Neelakantan, T. Xu, R. Puri *et al.*, 'Text and Code Embeddings by Contrastive Pre-Training', 24 January 2022. arXiv: 2201.10005 [cs]. [Online]. Available: http://arxiv.org/abs/2201.10005 (visited on 30/01/2022).

A. Processing of the BERT hidden layers

As input, the model accepts the index numbers of tokens from the earlier tokenization step. The resulting output of the model hidden states has the shape

$$13 \text{ layers} \times 1 \text{ batch} \times \text{number of tokens in word} \times 768 \text{ hidden units}$$

Since we are dealing with individual sentences, the batches dimension always equals 1 and is discarded. The layers and tokens dimensions are swapped to obtain the embeddings for individual tokens, resulting in a $\text{hidden states}_{\text{number of tokens} \times 13 \times 768}$ matrix for each sentence.

B. Text cleaning

The following regular expressions were used in the initial text cleaning step. Commands 1-2 remove hyperlinks. Command 3 removes text in parentheses and commands 4-5 removes asterisks, square brackets and underscores. Commands 6-7 remove certain Unicode symbols of unknown purpose found in some articles.

```
1 (\(http.*?\))
2 (\(%20http.*?\))
3 (\(.*?\))
4 \*|\[|\]
5 r'__'
6 u'\u200b'
7 u'\u2028'
```

For example, the original raw text

```
The case was **discussed** in [a document found online](http://example.com).
```

intended to be displayed on the web as

The case was **discussed** in a document found online.

results in

The case was discussed in a document found online.

after preprocessing.

C. The graphical user interface

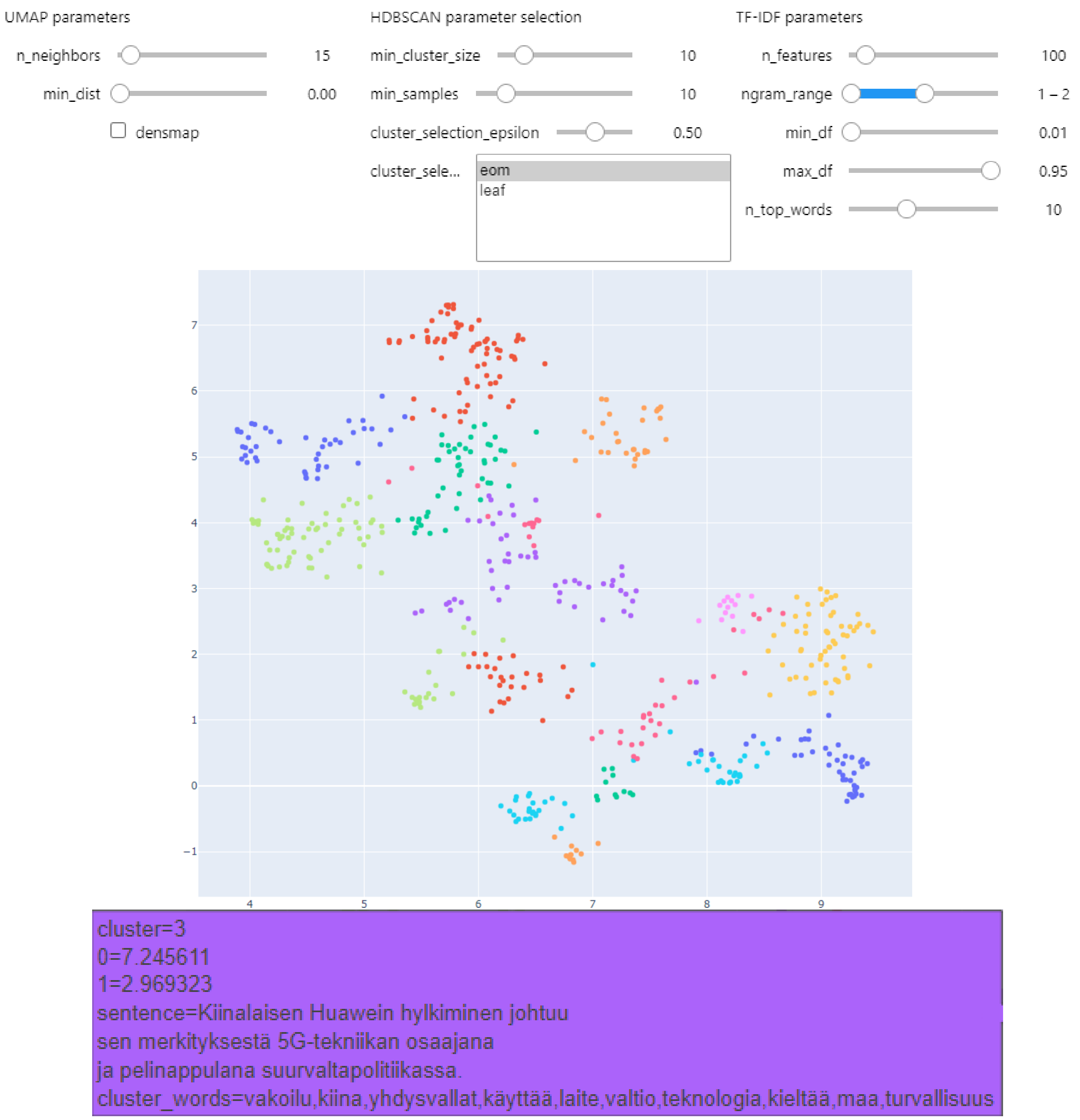


Figure 3.1. Results from changing the parameters (top) are presented in real time. A scatter visualization(middle), showing the sentence embeddings reduced to two dimensions in UMAP and colored labels corresponding to clusters. By hovering the mouse over individual data points, more information appears in an infobox (bottom).

D. Audit trail for thematic coding

D.1 Processing cluster keywords to initial codes

Keywords for cluster 0

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
5G	5G	0,36	onnistua	succeed	0,5	oikea	correct	0,5
turvallisuus	security	0,32	voittaa	win	0,33	omia	own	0,5
kauppa	trade	0,22	sulkea	close	0,27	valmis	ready	0,5
verkka	-	0,21	liikkua	move	0,19	mahdollinen	possible	0,43
ostaja	buyer	0,21	avata	open	0,19	uusi	new	0,27
kehite	development	0,21	palauttaa	return	0,19			
panos	invesment	0,21	luottaa	trust	0,19			
kohtelu	treatment	0,21	huolestua	worry	0,19			
juna	train	0,21	kannattaa	support	0,19			
markkinat	markets	0,19	neuvotella	negotiate	0,17			

No sensible coding was deemed possible for this cluster.

Keywords for cluster 1

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
johtaja	director	0,41	rauhoitella	calm	0,39	luonnollinen	natural	0,39
huolestunut	worried	0,33	pidättää	arrest	0,34	laiton	illegal	0,39
pidätys	arrest	0,33	rakentaa	build	0,27	kansallinen	national	0,39
luovutus	extradition	0,28	lähteä	leave	0,26	ahdas	tight	0,39
kumppani	partner	0,23	kadota	disappear	0,26	avoin	open	0,34
5G	5G	0,19	antaa	allow	0,23	oma	own	0,3
toimi	job	0,19	pelätä	fear	0,23	paha	bad	0,3
siru	chip	0,17	lopettaa	quit	0,2	hyvä	good	0,25
sovellus	application	0,15	sulkea	close	0,18	uusi	new	0,21
verkko	network	0,13	kieltää	ban	0,18			

Two codes were generated for cluster 1 keywords. The first one, consisting of the verbs [*disappear, end / terminate, break / sever*] and the noun [*ties*] was coded together on the basis of a top-level concept named **cessation, termination, ending**.

The second code was named **distress or trouble**, based on the verbs [*/to* *be] hard-pressed*] and [*to limp*].

Keywords for cluster 2

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
päivitys	update	0,6	hyötyä	benefit	0,32	uusi	new	0,59
puhelin	phone	0,36	käydä	visit	0,32	tekninen	technical	0,36
muutos	change	0,21	saada	receive	0,3	vaikea	difficult	0,36
toiminta	activity	0,19	antaa	give	0,28	ehdoton	absolute	0,36
myynti	sales	0,17	jäädä	stay	0,24	varma	certain	0,36
käyttö	usage	0,15	sanoa	say	0,17	todellinen	real	0,36
aika	time	0,14	esiintyä	present	0,16			
kauppa	trade	0,13	näkyä	be visible	0,16			
into	excitement	0,12	puolustaa	defend	0,16			
luottamus	trust	0,12	raottaa	reveal	0,16			

The words [*to upgrade / update, phone*] and [*technical*] were coded as **devices and technical matters**. The second code, consisting of the verbs [*to benefit, to receive, to give*] was more abstractly coded as **mutually beneficial activities**.

Keywords for cluster 3

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
puhelin	phone	0,35	taittua	fold	0,49	oma	own	0,46
kännykkä	phone	0,29	julkaista	release	0,42	uusi	new	0,32
näyttö	screen	0,24	tulla	come	0,23	iso	large	0,31
operaattori	service provider	0,21	toimia	work	0,23	sileä	smooth	0,2
järjestelmä	system	0,21	saada	get	0,21	tyhjä	empty	0,2
tekniikka	technology	0,18	kertoa	tell	0,19	tieto	knowledge	0,2
käyttö	usage	0,18	käyttää	use	0,19	eri	different	0,2
teknologia	technology	0,15	testata	test	0,16	turha	useless	0,2
yhtiö	company	0,14	hoitaa	take care of	0,16	länsi	western	0,2
yllätys	surprise	0,14	aueta	open	0,16	kallis	expensive	0,2

Two codes were identified, the first one was named **mobile phones** based on its keywords [*phone, mobile phone, display (of a device), service provider, to release [a product]*]. The second one, **assessing quality**, included the words [*unnecessary / useless, expensive, to test*].

Keywords for cluster 4

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
huoli	worry	0,46	lobbata	lobby	0,29	optimistinen	optimistic	0,43
toiminta	activity	0,29	rajoittua	limit	0,29	varovainen	careful	0,43
kauppa	trade	0,29	tarkkailla	supervise	0,29	ymmärrettävä	understandable	0,43
haastattelu	interview	0,23	viitata	refer	0,25	säännöllinen	regular	0,43
väli	interim	0,23	onnistua	succeed	0,25	mahdollinen	possible	0,37
syytös	accusation	0,23	herättää	wake	0,25	erityinen	special	0,37
sana	word	0,23	arvioida	assess	0,25			
viestintä	communications	0,23	uskoa	believe	0,22			
etu	benefit	0,23	vakoilla	spy	0,22			
raportti	report	0,23	seurata	follow	0,22			

The coding for this cluster was less straightforward than the earlier ones. While several interpretations and judgement calls were possible, a degree of arbitrariness was inevitable in deciding between them. All of the four final codes were related to abstract activities, communications or both. The coding was influenced by an assessment of emotional valence.

The code **communications and public messaging** ([*word, messaging/communications, report*]) was neutral and more clearly related to communications.

With both communicative and thinking-related interpretations being plausible, the words [*worry, accusation*] were coded together as **voicing concerns and accusations**.

cautiously assessing an evolving situation ([*to assess, to monitor, to follow/to keep an eye on*]) was considered separate from **expressing understanding** ([*optimistic, careful/cautious, understandable*]), with the former denoting activity and the latter internal states.

Keywords for cluster 5

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
ratkaisu	solution	0,41	saada	get	0,48	vaarallinen	dangerous	0,37
verkko	network	0,39	ostaa	buy	0,35	yhteinen	common	0,37
laite	device	0,31	sanoa	say	0,27	kriittisä	critical	0,37
aika	time	0,28	hakeutua	gravitate	0,25	laadukas	high-quality	0,37
asema	position	0,24	panostaa	invest	0,25	viides	fifth	0,37
laatu	quality	0,24	operoida	operate	0,25	kaksi	two	0,37
panostus	investment	0,24	lyödä	strike	0,25	vahva	strong	0,28
ekosysteemi	ecosystem	0,24	kiinnostaa	interest	0,25	hyvä	good	0,23
markkinat	markets	0,22	toivoa	hope	0,22	suuri	large	0,22
palvelu	service	0,18	viitata	refer	0,22			

One code was named **markets and their interconnectedness**, with [*the market*] as a top-level concept and [*ecosystem*] as a qualifier.

The second code (**emphasising quality**) was closely related to the first but considered separate based on emphasising actions ([*an investment, to invest in*]) as well as judgement calls ([*quality, good-quality, strong, good*]).

Keywords for cluster 6

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
syytös	accusation	0,57	kiistää	deny	0,91	tällainen	like this	0,69
väite	claim	0,41	kommentoida	comment	0,22	kaikenlainen	all sorts of	0,4
vakoilu	spying	0,36	torjua	reject	0,14	tavallinen	regular	0,4
asia	issue	0,29	kieltää	deny	0,11	ensimmäinen	first	0,34
epäily	suspicion	0,24	linkittää	link	0,08	koko	entire	0,31
sivu	side	0,19	tyrmätä	knock out	0,08			
yritys	company	0,13	syrkiä	discriminate	0,08			
kauppa	business	0,1	esittää	propose	0,08			
johtaja	director	0,1	ratkaista	solve	0,08			
dokumentti	document	0,09	poistaa	remove	0,08			

voicing concerns and accusations: [*accusation, claim, suspiction*]

vehement denial: [*to deny, to rebut, to reject*]

expressing opinions: [*to comment*]

Keywords for cluster 7

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
ykkönen	number one	0,38	nousta	rise	0,6	suuri	large	0,8
kärki	peak	0,33	menestyä	succeed	0,31	kolmas	third	0,3
markkinat	markets	0,31	hengittää	breathe	0,31	vakava	serious	0,17
niska	neck	0,28	kasvaa	grow	0,22	nopea	fast	0,17
valmistaja	manufacturer	0,28	tulla	come	0,22	yli	over	0,17
puhelin	phone	0,24	kilpailla	compete	0,16	nope	fast	0,17
verkko	network	0,23	tavoitella	strive	0,16	poikkeuksellinen	unusual	0,17
laite	device	0,2	verrata	compare	0,16	koko	size	0,13
tulos	result	0,19	kuulua	be heard	0,16	iso	big	0,13
kilpailija	competitor	0,18	huohottaa	breathe	0,16	vahva	strong	0,13

success in competition and growth: [*number one the top to rise to grow to succeed*] **markets and competition**: [*competitor to compete the market manufacturer profit*] **metaphorically taking over in a competitive situation**: [*to breathe to gasp neck*]

Keywords for cluster 8

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
5G	5G	0,48	valmistaa	manufacture	0,61	edistynyt	advanced	0,68
verkko	network	0,33	maksaa	pay	0,44	halpa	cheap	0,68
operaattori	service provider	0,32	sijaita	be located in	0,38	toinen	other	0,26
verkka	network	0,31	asentaa	install	0,34			
markkinat	markets	0,29	toimia	operate	0,3			
laite	device	0,27	rakentaa	build	0,3			
hetki	moment	0,26						
laita	device	0,21						
keskittymä	conglomeration	0,21						
kännykkö	phone	0,16						

No sensible coding was deemed possible for this cluster.

Keywords for cluster 9

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
kehitys	research	0,5	perustaa	found	0,6	suuri	large	0,45
tutkimus	development	0,5	kasvaa	grow	0,32	mobiili	mobile	0,38
tuote	product	0,32	palkata	hire	0,23	pohjoinen	northern	0,38
maailma	world	0,32	harjoittaa	practise	0,23	myyty	sold	0,38
keskus	center	0,31	työllistää	employ	0,23	johtava	leading	0,33
viestintä	communications	0,16	siirtyä	transition	0,23	kolmas	third	0,33
puhelin	phone	0,14	rekrytoida	recruit	0,23	toinen	second	0,3
verkka	network	0,09	investoida	invest	0,23	uusi	new	0,21
arvo	value	0,09	lupailla	promise	0,23			
tehtävä	task	0,09	hallita	govern	0,2			

research and development: [*development research*] **the job market**: [*to hire to employ*] **economic benefits**: [*to invest in*]

Keywords for cluster 10

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
5G	5G	0,62	rakentaa	build	0,45	usea	several	0,67
maa	country	0,48	kieltää	forbid, ban	0,45	merenalainen	undersea	0,44
verkko	network	0,31	estää	prevent, block	0,30	ensimmäinen	first	0,38
rakentaminen	construction	0,26	päättää	decide	0,26	vastainen	opposed	0,30
käyttäminen	usage	0,23	päästää	let, allow	0,26	suuri	large	0,26
tekniikka	technology	0,19	harkita	consider	0,22	uusi	new	0,24
markkina	market	0,10	pohtia	deliberate, entertain	0,20			
verkka	NA (mislemmatized)	0,08	sulkea	shut down	0,18			
yhtiö	company	0,08	ilmoittaa	announce	0,18			
muu	other	0,08	sanoa	say	0,14			

Keywords for cluster 11

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
oikeus	court	0,91	haastaa	challenge	0,88	syytön	innocent	0,82
päätös	verdict	0,21	hylätä	abandon	0,18	vastainen	opposed to	0,57
ministeriö	ministry	0,15	valittaa	complain	0,18			
kanne	claim	0,15	jättää	leave	0,18			
varastaminen	stealing	0,15	voittaa	win	0,15			
rikkominen	breaking	0,13	todeta	establish	0,15			
pakote	sanction	0,12	hävitä	lose	0,14			
tapaus	case	0,12	koskea	touch	0,12			
kiista	dispute	0,1	syyttää	accuse	0,12			
valmistaja	manufacturer	0,09	tulla	come	0,12			

legal proceedings: [*court of justice lawsuit to sue to prosecute not guilty*]

Keywords for cluster 12

Nouns			Verbs			Adjectives		
Word	Translation	Weight	Word	Translation	Weight	Word	Translation	Weight
vakoilu	spying	0,69	epäillä	suspect	0,65	usea	several	0,49
laite	device	0,35	käyttää	use	0,44	vastainen	opposed	0,44
turvallisuus	security	0,22	rikkoa	break	0,28	suuri	big	0,37
maa	country	0,2	varoittaa	warn	0,24	rankka	harsh	0,31
riski	risk	0,17	syyttää	blame	0,2	musta	black	0,31
verkko	network	0,15	hyödyntää	exploit	0,19	kireä	tight	0,31
mieli	mind	0,15	pitää	must	0,16	salaisin	classified	0,31
huoli	worry	0,13	auttaa	help	0,09	hyvä	good	0,2
länsi	west	0,13	painostaa	pressure	0,09			
tapaus	case	0,11	väittää	claim	0,09			

expressing suspicion: [*to suspect, to claim*] **expressing accusations**: [*to accuse*] **expressing concern**: [*worry*] **being in opposition to**: [*opposed to*] **warning and pressuring others to take action**: [*to warn, to pressure*] **geopolitical significance**: [*country the West*]