Linköping University, Department of Computer Science
729G81, Applied Cognitive Science

# Identifying and Analysing Misinformation about The National Board of Health and Welfare

## A Data-driven Approach using Web Scraping, Sentiment Analysis, and Topic Modelling

Adrian M. Brynielsson, Artur Ladjarov, Joel K.O. Moche, Marcus D. Grattan, Olle F. Glöde & Victor A. Karlströmer

May 24, 2023

*Supervisors: Fredrik P. Stjernberg, Robin T. Keskisärkkä*

*Examiner: Arne N. Jönsson*

LINKÖPINGS UNIVERSITET

*This page intentionally left blank*

Abstract

In the age of digital information, the proliferation of misinformation has become a pervasive and substantial issue. The social implications of such misinformation include distorting democratic processes and undermining trust in public institutions. This project's purpose is to provide KAPI with valuable data concerning potential online misinformation narratives about the role of Swedish social services (Socialtjänsten) in child welfare, leveraging information from a variety of social media platforms. By harnessing a data-driven approach that integrates methods such as web scraping, sentiment analysis (using VADER), and topic modelling (using BERTopic), we aim to extract, analyse, and categorise online discourse on this subject. The main objective is to better understand the prominent narratives and sentiments related to Swedish social services and to identify the main channels through which misinformation is disseminated. This project provided seven primary potential misinformation narratives consistently prevalent across the discourse, each offering distinct perspectives with often negative sentiment towards Swedish social services. The implications of our findings can provide insights to policymakers, researchers, and public administration officials, contributing to a better understanding of how misinformation spreads.

## Acknowledgements

We express our sincere thanks to KAPI for their indispensable support throughout this project, including proofreading, providing essential materials, and continuous guidance. We also extend our gratitude to our supervisors, Robin Keskisärkkä and Fredrik Stjernberg, for their expert advice and direction.

We extend our gratitude to Arne Jönsson, who served as our examiner, ensuring the project's academic rigour. Acknowledgement is also due to the Swedish National Board of Health and Welfare, Socialstyrelsen, for indirectly providing the opportunity for our involvement.

We additionally value Försvarshögskolan for their insightful reports, which were instrumental in shaping our understanding of the topic at hand. Lastly, we recognise our university for providing the conducive environment and resources necessary for our research. The success of this project would not have been possible without the collaborative efforts and support received, for which we are deeply grateful.

# Table of Contents

# 1. Introduction

In today's digital age, social media platforms play an integral role in shaping societal perceptions and opinions. The rapid exchange of ideas and information that these platforms enable is a double-edged sword. While facilitating dialogue and fostering connectivity, they simultaneously provide an ideal platform for the spread of misinformation. This misinformation, spreading swiftly and broadly, has the potential to distort public sentiment and understanding, necessitating rigorous study and understanding to curb its societal impact.

The focus of our research stems from a project commissioned by KAPI for the Swedish National Board of Health and Welfare (from now, Socialstyrelsen). In 2023, the government instructed Socialstyrelsen to scrutinise and report on the myths and rumours circulating about the Swedish social service (from now, Socialtjänsten) work concerning children and families. The ultimate goal is to monitor and compile ongoing misinformation to gain deeper insights that would help shape the future communication strategies of Socialstyrelsen.

In line with this objective, our project specifically concentrates on identifying and analysing misinformation about Socialtjänsten. Utilising a data-driven approach involving web scraping, sentiment analysis, and topic modelling, we aim to examine the dynamics of this misinformation within a collected dataset of ~400 000 social media posts and comments from various platforms.

The subsequent sections of this paper will provide a comprehensive overview of our project's background, the theoretical framework guiding our approach, and the methods employed in data collection and analysis. We will present the findings from our analysis, discuss their implications, and conclude with a summary of our insights and suggestions for future work.

By engaging in this exploration, we seek to contribute to both academic discourse on the subject of misinformation on social media and provide valuable insights that could inform the creation of strategies to counteract misinformation. The overarching goal is to facilitate a more transparent and well-informed public dialogue about the role and work of Socialstyrelsen and similar institutions.

# 2. Background

Misinformation is the spreading of intentionally false information through media or other public channels. This could be anything from hoaxes, propaganda or false narratives being presented as facts. Considering that this project revolves around fake news, how do we define it? “Fake news” is a fairly recent term used synonymously with misinformation. The reason for the term “fake news” having been used so frequently owes up to the 2016 presidential election in the United states. According to Google Trends[1], a publicly accessible tool to compare search terms frequency after time, searching “fake news”, you can see a clear difference in September of 2016. The term goes from virtually un-searched to very popular in conjunction with the 2016 presidential campaign. The term was however popularised due to politicians using it as a way of denying reports that would discredit the party or their individual members. For example, a New York Times article citing a release of Donald

[1] Google trends (n.d) “*Fake news”,* Retrieved April 29, 2023. https://trends.google.com/trends/explore?date=all&q=%22fake%20news%22

Trump's tax returns from 2005 was met by a tweet from Donald Trump himself stating that the papers were falsified and labelling it fake news (The New York Times, 2017).


*Fig 1. Google Trends results over time since 2004. The graph shows the term "fake news" and the amount of searches that have been made on Google. Y-axis is the amount of search relative to the total amount of Google searches.*

Despite the recent popularisation of the term, misinformation has long been a tool to manage influence or control opinion over a certain population, and throughout history it has been utilised in different ways. With an ever-changing media landscape due to advanced technology, what used to be a tool employed in traditional media, has instead found its way to social media and public internet forums. And with the increased strength of AI and the possibility of anonymity, malicious actors can sway opinion without being held accountable. Darren Linvill (2019) explains that Russian Twitter bots belonging to the Internet Research Agency (IRA) promoted hostile conversations revolving around anti-vax sentiments disguised as real Twitter users, in an attempt to increase vaccination hesitancy.

So what is the purpose of fake news? What are the concrete consequences of a deliberate campaign of misinformation? Antagonistic or foreign actors use misinformation as a way of indirectly accomplishing their own purposes and goals in a target nation (Ranstorp & Ahlerup, 2023). Misinformation today has shifted from a model where the goal was to form narratives and create stories that were constructed to the benefit of its creators, to something more complex. Considering the vast amounts of information that is available for free, creating false narratives simply is no longer a viable option. Instead, the misinformation works more intensively by questioning the truth while it has not been proven, with the goal of sowing distrust for the democratic society and structure, or to launch hostile discussions intended to divide the population (McKay & Tenove, 2021).

According to a report released by the Swedish Defence University (Ranstorp & Ahlerup, 2023), misinformation and its components can be divided into many categories. Modern misinformation is based on "storytelling" and the narrative of this storytelling can be summed up into three topics: narratives regarding the current international order, narratives with a purpose of influencing aspects of policy, and narratives circulating identity. All three narratives are highly relevant in the modern political atmosphere but can interpersonally be very sensitive topics to the individual citizen. This makes them valuable weapons in sowing distrust or creating division amongst populations. Several narratives can be used in conjunction with each other, which enhances their effectiveness by making the individual narratives seem more convincing through the creation of larger overarching narratives. Misinformation can also be divided into constructive, discrediting and evasive narratives. To begin with, constructive narratives are trying to construct, complete or evolve already existing narratives. Discrediting narratives exist to eliminate or undermine already existing or future narratives. Evasive narratives are used to create confusion or a distraction in relation to an already existing narrative. The report also defines and explains the words “campaign” in the context of fake news. A campaign consists of several elements, working in parallel but with different objectives, targets and methods, but in the end intends to affect the same narratives and motives of its stakeholders.

But is misinformation actually a viable strategy to change narratives? Considering social media is a relatively new medium of communication, there is not huge amounts of literature investigating results in modern fake news campaigns. But there are still reports indicating that it works. One of the most famous campaigns in the last decade was in preparation for the 2016 United States presidential election. In July of 2018, the FBI issued a statement returning the indictment of 12 Russian intelligence officers interfering with the 2016 election (FBI, 2018). These intelligence officers were charged with conspiracy to hack computers of U.S persons and entities involved in the election, stealing documents and later releasing them, all in conspiracy of altering opinions and the outcome of the election. With Russian intention to interfere being proven on public forums, Stachofsky et.al (2023) researched news consumption, voting methods and trust of election processes for the 2020 presidential election. In their paper, they found a significant effect on people who consumed hyper-partisan or fake news on their disbelief in the election process. This was propagated through, mainly facebook and alternative media, being a source of enabling the spread of desinformation through the utilisation of fake news. This combined with the proof of Russian interference shows that efforts in controlling the narrative has a significant effect on the intended target.

When initiating a fake news campaign, you want to involve the topics that are relevant to the target audience and are current with news and media. The foreign actor involved in this campaign that we are basing this paper on, has targeted a majority muslim minority in Sweden, with origins from north and east africa, as well as the middle east and western parts of asia. The reason for this is the conflict between Swedish culture, lifestyle and society, and muslim tradition (Ranstorp & Ahlerup, 2023). This cultural difference has on several occasions flared up into actual conflict, where debates and even demonstrations have escalated polarisation of opinion. This effect is exactly what this antagonist is looking to gain from these campaigns (Ranstorp & Ahlerup, 2023). Direct violence and insurgency is not a viable option to destabilise a country, while polarisation and growing distrust of government and authority is a way to indirectly create violence (Fedorenko & Fedorenko, 2022). One such example is the planned burnings of the Quran by Rasmus Paludan in 2022, where violence flared up by Swedish muslims being angered by what is conceived as a blasphemous act. The government's choice of protecting Swedish free speech laws was received poorly and reduced trust in the Swedish legal system within the muslim minority (Ranstorp & Ahlerup, 2023). In an act to further reduce trust in Swedish authority, malicious actors have decided to target Socialtjänsten. Reporting from several news stations began in early 2022 of a fake news campaign creating rumours of Socialtjänsten kidnapping children of muslim background and forcibly placing them in christian homes (SVT, 2022; Dagens Samhälle, 2022). The foreign propagators fueling these rumours are, according to the Swedish Defence University, social media influencers (Ranstorp & Ahlerup, 2023).

A popular term to use by these propagators in relation to the fake news is the Swedish Care of Young Persons Act, shortened to LVU in Swedish. While it is true that Socialtjänsten can forcibly relocate children to new homes, several criteria have to be matched, and it is seen as a last resort measure (Socialtjänsten, 2020). However, due to the severity of the action and the large amount of laws and rules surrounding the act, it’s easy to make misinformed assumptions regarding LVU, and is therefore an easy target for criticism when discussing particular cases that come into media's attention. This also makes the act a prime tool for fake news stories, as an already delicate topic such as LVU can easily agitate an audience.

## 2.1. Project Background

This part of the paper will discuss the background and formalities of the project from our clients. It will reference reports from parties involved in the project that have been shared with everyone involved. Due to confidentiality, we have been allowed to reference the contents of these reports, but not provide any copies of these in the reference list. Everything that we have referenced from these reports have also been proofread by KAPI so as to not misrepresent them.

Due to the fake news campaign, known as the LVU-campaign, against Socialtjänsten and the possible effects of distrust for a governmental organisation, the Swedish government decided to act. Based on intelligence from the Swedish Secret Service (from now on, SÄPO), the government ordered an investigation of the fake news campaign against Socialtjänsten. This was to be done by Socialstyrelsen, which is the body for rules, statistics, education and evolution of Socialtjänsten. Socialstyrelsen offered a procurement to a communication bureau named KAPI. The procurement stated that KAPI was to perform a status report on the current discourse on social media, in five different languages. These reports are continuously provided until the end of 2023, and with them, KAPI presents an analysis on the data that they have collected. The goal with all of this work is to provide Socialstyrelsen with guidance of communication, both in the short term with the ongoing campaign, but also with a long term perspective in preparation of possible campaigns or controversies.

At the time of writing this report, KAPI has presented several status reports to Socialtjänsten. These reports consist of brief explanations of activity within several relevant languages and what has been said on social media that misrepresents Socialtjänsten or is not considered true. In alignment with guidelines set by Socialtjänsten, both we and KAPI only disclose the substance of the current discourse and conscientiously refrain from reporting who specifically said what. So far, according to the documents we have access to, the coverage has been done on Twitter, Facebook, Instagram, TikTok and Youtube. The languages that have been covered are Swedish, Somalian, Polish, Russian/Ukrainian, Arabic, Persian, and Dari. Notable observations are found in Somalian, Polish, Arabic, Persian and Dari, where it is reported that many posts are revolving around Socialtjänstens relocation of children, and LVU, and hold a strong scepticism, perturbation and generally contain a negative sentiment. Involved in this project is the Psychological Defence Agency (from now, MPF). MPF provides reports on their own in parallel with KAPI, but focuses on foreign actors and trends in activity when it comes to fake news. For example, on 27 of March, MPF reported a small increase in narratives regarding Socialtjänsten and LVU, although from the same few actors, but on 21 of April, this increase was reduced to its normal rate of activity. This small spike in activity was probably an effect from a recent case where an elderly man was found dead in his jail cell after being arrested for child abduction of his own daughter. Important to note is that his passing was due to illness, and no foul play is suspected (Expressen, 2023; DN, 2023).

Considering that KAPI is working with this until the end of 2023, we were given freedom to organise our work after our deadlines and not KAPI's. Consider the goals as this paper's research questions. What KAPI wanted from us was to create a scraper that could extract data from social media posts. With this data, we were to do an analysis of our own choice that we saw fit for a paper in cognitive science, but also of value for KAPI. We were given instructions of platforms that were of interest and a list of words that we were to use to filter the data scraped from these platforms. With weekly meetings the parties informed each other on updates or other things that were of value. The goal with the analysis was to find narratives in what was being said. Using a data driven approach, can we prove the same narratives that are being reported from KAPI and Socialtjänsten instead of using qualitative

# 3. Theory

The following section will provide a detailed account of the theoretical and technical foundations for each of the methods used within our project.

## 3.1. Web Scraping

There are several ways to extract data from the internet, amongst these are usage of Application Programming Interfaces (API). An API is a set of protocols and tools that allows for the retrieval of data from a web application. Many social media companies such as Facebook, Instagram and Twitter provide API services that allow access to their users' data.

These services are often limited in what type of data one can extract, the amount of data you access as well as being monetized. Furthermore, the usage of company provided APIs places limits upon the user with regards to what the data can be used for. Therefore other methods for extracting data from websites might be more useful and cost-effective.

Web scraping is a method used to extract information from websites and convert it into easily manageable formats such as CSV, databases, or JSON. The process of web scraping can however be a very resource and time consuming task, especially when done manually. Therefore techniques and tools to automate the process are necessary in order to make it suitable for extracting large amounts of data. Thus a web scraper is a type of software that can mimic a human browsing the internet and gather information from web-documents. The benefit of doing this with web scraping rather than manually doing it by hand is that a scraper has the ability to collect this information quickly and automatically. This also facilitates the amount of data that is scraped, where you could scrape, for example, hundreds of thousands of posts in a relatively short amount of time (Diouf, et al., 2019).

Since websites are most often composed of documents in HTML-format the extraction of information from these documents involves parsing such files. Common web scraping methods make use of the inherent structure of HTML documents to extract certain elements from the document. HTML-documents have a hierarchical structure, which can be represented as a tree-like structure known as the Document Object Model (DOM). The DOM represents the HTML elements and their relationships as a tree, where relationships between elements can be represented as parent-child or sibling relationships. Web scraping leverages this tree structure to navigate and extract specific elements from the document.

One commonly used technique used when web scraping is the construction of so-called XPaths described using the XML Path Language. These allow web scrapers to specify paths in the tree structure used to select elements based on their type or other meta attributes and extract the information contained within these elements.

There are two main ways of constructing web scrapers. Some scrapers utilise browser automation tools, which allow the user to interface with a web-browser through user-specified code. These tools simulate user interaction, such as clicking buttons and scrolling. Such tools are commonly used when scraping dynamically loaded websites as these allow the scraper to handle dynamic elements such as dropdown-boxed and waiting for elements to load. Furthermore, because a browser is used the programmer does not need knowledge of the underlying protocols such as HTTP (HyperText Transfer Protocol) and/or TCP (Transport Control Protocol) which other methods require. Thus, the scraper can bypass the complexities of the underlying network communications, allowing for easier navigation and handling of complex web-flows.

The main drawback of such methods, however, is the execution speed. As these processes involve rendering web-pages executing JavaScript code, the process is considerably slower compared to other methods.

On the other hand, web scraping can be performed by directly communicating with the desired host server via sending HTTP requests and parsing the HTML responses. Such scrapers, commonly referred to as “headless” or “non-browser” scrapers, work at a lower level, requiring knowledge of the underlying network protocols. Thus, to properly communicate with end-servers it is required of the web scraper to properly manage application-layer parameters such as manually constructing headers for HTTP communication, handling authentication mechanisms and managing sessions to name a few. However, this approach has the advantage of faster and more lightweight scraping solutions as they eliminate the overhead of browser rendering.

When performing web scraping it is important to consider the legal and ethical aspects around it. There are several factors to consider, including the website being scraped, the laws surrounding web scraping, and the purpose of the scraping. In general, web scraping of publicly available data is legal from websites, as long as it does not infringe on any trademark, copyright or intellectual property rights. However, it is important to be aware that some websites have terms of service or usage policies that might prohibit web scraping, if you would scrape these websites anyway it might result in legal consequences. Web scraping that involves accessing data that is password-protected, private or by bypassing security measures to gain access to a website is in general illegal and could lead to criminal charges. This kind of scraping violates security and privacy laws which may result in financial and legal penalties. Therefore it is important to ensure that web scraping is done in compliance with these regulations and laws and respect the terms of service to avoid any legal issues. Another important factor to consider is that even though web scraping might be technically legal it could still raise ethical concerns, particularly when it involves scraping personal data (Krotov & Silva, 2018). To ensure compliance with the law our preprocessing step anonymizes the data.

## 3.2. Data Pre-Processing

After data has been extracted and collected, the next step in the process of data analysis is to clean the data. This is a critical step in ensuring that the results of data analysis are reliable and valid. It involves identifying and correcting errors, inaccuracies and inconsistencies in the data to ensure that data is complete and relevant for the analysis. Another important factor of data cleaning is correctly formatting the data, meaning that it should be consistent and formatted suitably for analysis. This could involve removing potential HTML code and standardising the data format, such as ensuring that the dates are in the same format. For example, if data has been scraped from several different websites this might result in different formats in the output, so having the cells formatted consistently over all CSV files is essential in order to make the analysis straightforward. Also, we need to consider removing certain types of data and duplicates. When web scraping social media posts, usage of emojis are common in the data, this could complicate the analysis since the meaning of these emojis might differ which makes it hard to interpret and could lead to an incorrect analysis. Furthermore, while performing subsequent steps in the analysis the presence of emojis could cause errors in the text processing. Additionally, in our Swedish-language data set, the presence of non-ascii characters and special characters or symbols are also factors that can negatively impact the analysis. By removing these inconsistencies it can help improve the relevancy of the data as well as the accuracy of the analysis (Zou, 2022).

Lemmatization can also be considered when performing data pre-processing. Lemmatization is the process of reducing a word to its base root form, which is called a lemma. For example the word “jumping” can be reduced to its base form “jump”. The reason to do this is to group words with the same meaning and reduce the number of unique words in the dataset. This is useful in natural language processing applications, such as topic modelling or sentiment analysis since it enables an easier process of identifying relationships and patterns in the data (Pramana, et al., 2022).

## 3.3. VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based model designed for sentiment analysis of social media text developed by C.J. Hutto & Eric Gilbert (2014). The model focuses on analysing sentiment of informal and short texts which is common on social media platforms like Facebook, Twitter and Flashback. It works by utilising a pre-built sentiment lexicon with words that have been assigned sentiment scores. The sentiment scores indicate the sentiment of the word, basically how positive or negative a word is. The lexicon also includes intensity modifiers to capture the degree of sentiment intensity. VADER takes the context of the words into account by considering the words that follow or precede them, as well as calculating an intensity score for each word to capture the level of sentiment expressed. By looking at preceding words and words that follow it also take intensifiers and diminishers in the text into account to adjust the intensity score, for example, words like “incredibly” or “highly” can increase the intensity, while words like “somewhat” can decrease it. After VADER has analysed the individual words the model combines the scores to calculate a sentiment polarity score for the whole sentence. The polarity score is determined by summing up the individual word scores while considering the sentiment intensification and grammatical rules. VADER also incorporates a scoring and mapping for emoticons, slang and acronyms which is common in social media.

## 3.4. Topic Modelling

Topic modelling is a method for discovering underlying themes or topics in a collection of language data. The goal of topic modelling is to identify latent topics present in the data, and give a representation of the data as a mixture of these topics. This method is particularly useful when working with large datasets and manual analysis is impractical or infeasible. Topic modelling does not require predefined categories or labels and therefore enables discovery of themes not previously considered since it allows unbiased exploration of the data. Topic modelling is useful in various different domains and industries, such as social media analysis, market research, content recommendation and more. By grouping overall opinions into topics, organisations can make informed decisions on emerging trends and better understand the different narratives being spread. The method does have its limitations, the quality of the topics are heavily dependent on the quality of the data and the preprocessing steps such as data cleaning. Nonetheless, when topic modelling is used appropriately it can be a powerful tool for analysis and knowledge discovery in large datasets (Vayansky & Kumar 2020).

There are several different methods for topic modelling, amongst these methods are Latent Dirichlet Allocation (LDA) and BERTopic. LDA, which arguably is the most popular topic modelling technique, is a generative probabilistic model of a corpus. It is based around the idea that each document in the corpus is seen as a combination of different topics and each topic is characterised by a distribution of words. Within each topic the words with the highest probabilities generally give a good representation of what the topic is (Jelodar et al., 2019).

It has been found however that LDA neglects the syntactic links among words since it does not take semantic relations into consideration of words within a phrase. It rather presents the topics as a bag-of words. To solve this problem text embedding approaches have emerged as a solution (Udupa et al., 2022).

### 3.4.1. Transformer Embedding Models

Transformer embedding refers to the process of encoding text to dense vector embeddings. Dense vector embeddings are used to numerically represent semantic meaning by encoding abstract meaning and relationship of words. Dense vectors tend to be highly dimensional, with the 784-dimensional approach being common. Each dimension carries relevant information as determined by a neural network. There are many ways to generate these embeddings, with four common ones being word2vec, sentence transformers, and dense passage retrievers (DPR). Of the aforementioned approaches we used sentence transformers within our application.

### 3.4.2. Sentence Transformers

Sentence transformers produce information rich dense vectors, with a variety of use-cases such as sentiment analysis and question answering. Dense vectors are backed by a double array representing its entry value while sparse vectors are backed by two parallel arrays, indices, and values. As a result a dense vector uses more memory but is also able to contain more information. Because of this, transformers are dominant in modern language models. BERT is one of the most used transformer architectures, but the following description applies to most transformer models. These work by producing vector embeddings for each token. This is similar to word2vec, but because of deeper networks the embeddings are more information rich, and because of a mechanism known as attention, the context of words is also encoded. The approach of embedding individual words is not optimal for long pieces of text as it limits the length of the text being embedded. Therefore sentence transformers then conduct a pooling operation to generate a sentence embedding from the individual token embeddings (Reimers & Gurevych, 2019). Within the category of sentence transformers, we specifically utilised a model called *Paraphrase-multilingual-mpnet-base-v2*. Given our use of this particular model, a detailed description of it will be provided.

The model works by mapping sentences and paragraphs to a 768 dimensional dense vector space, and is primarily useful for clustering and semantic search. It is built using a *Siamese BERT-network*. In this context, “Siamese” refers to a neural network containing two or more identical subnetworks, where the weight adjustments are mirrored within the models. These networks each generate an encoding for a sentence, where the cosine similarity between the encodings is then calculated and passed through an activation function, such as a sigmoid function, which outputs a similarity score (Reimers & Gurevych, 2019).

To train these Siamese convolutional neural networks SBERT uses a concept called triplet loss. Triplet loss functions by choosing a random anchor point in the data. For this anchor point it finds a positive pair and a negative pair, which refers to a datapoints belonging to the same class and a different class. These classes are based on euclidean distance, where the sentences with minimal euclidean distance are considered a pair. Afterwards the network calculates the similarity score for each of the two pairs. These 2 similarity scores are subsequently used to calculate a loss score, which is then used for backpropagation. Paraphrase-multilingual-mpnet-base-v2 is "multilingual" because it can process text in multiple languages as it has been trained on a wide range of them, which is partly why we chose to use it for our Swedish data (Reimers & Gurevych, 2019).

### 3.4.3. Dimensionality Reduction

Dimensionality reductions are used to reduce the dimensions of the data, this is done to allow for the next step, clustering, to be done. UMAP is a technique which keeps parts of the datasets local and global structure, this allows the creation of semantically similar clusters.

Uniform Manifold Approximation and Projection (UMAP) is a novel manifold learning technique used for dimensionality reduction. It is based on a theoretical framework within Riemannian geometry and algebraic topology. It has no computational restrictions on embedding dimensions, which makes it viable as a general purpose dimension reduction technique for machine learning. The uniform manifold refers to a technique that aims to preserve the local structure of data. It assumes that data points that are close to each other in the original data should also be represented as close to each other in the dimensionally reduced data, doing the same for data points which are further away from each other. Approximation and projection refers to the process of converting the original high dimensional data into lower-dimensional representations. UMAP is used to reduce dimensionality of the word embeddings, visualise the topics in lower-dimensional space, enhance interpretability and to more efficiently handle large-scale datasets (McInnes et al., 2018).

### 3.4.4. Clustering Models

After reducing our embeddings we are now able to cluster our data without being subject to the problem of clustering 768- dimensional data.

HDBSCAN is a density based clustering technique which can find clusters of different shapes, with the additional ability to identify outliers, which results in not forcing documents into clusters where they do not belong. This reduces noise. HDBSCAN works in five steps (Campello et al., 2013)

As described by Campello (2013) HDBSCAN works in the following way: Firstly, it transforms the vector space according to its density and sparsity. Secondly, it builds a minimum spanning tree of the weighted distance graph. Thirdly, it constructs a cluster hierarchy of connected components. Fourthly, it condenses the cluster hierarchy based on minimum cluster size. Finally, it extracts the stable clusters from the condensed tree.

### 3.4.5. Bag-of-Words

Given that clusters may possess different densities and shapes, a centroid-based topic representation technique, such as TF-IDF, which will be mentioned in a subsequent section, may not necessarily create the best fitting model. To circumvent these otherwise built-in assumptions, a “bag-of-words” approach is employed at the cluster level. This approach counts the frequency of every word within each cluster. Since we are interested in words on a topic level this allows us to find distinctions between clusters with otherwise similar words.

### 3.4.6. Topic representation using c-TF-IDF

Topic representation is a method used to distinguish one cluster from another. To accomplish this, we used a model known as Class-based Term Frequency Inverse Document Frequency (c-TF-IDF). Unlike traditional models that treat each cluster as a collection of individual documents, c-TF-IDF regards each cluster as a single entity. C-TF-IDF is a class based version of TF-IDF, which takes into account the context in which a word appears. c-TF-IDF works in several steps. First, Term Frequency (TF) for each document in the corpus is calculated. TF is the frequency of every word in every document. Inverse Document Frequency (IDF) is the measure of importance of a term in the entire corpus, it is calculated as the logarithm of the ratio of the total number of documents to the total number of documents which contains the term. IDF in essence gives higher weight to terms which are rare in the corpus but appear in specific clusters. Contextual Term Frequency (c-TF) takes into account the frequency of phrases/n-grams in the document instead of individual words . Contextual Inverse Document Frequency (c-IDF) measures the rarity of n-grams within the corpus, then uses the same formula as IDF but using these n-grams instead of the individual words. c-TF-IDF’s final calculation is done by multiplying the c-TF value of a term with the c-IDF value associated with the same term. The resulting score represents the importance of a term in the context of the document/class/cluster it belongs to and the importance within the entire corpus (Grootendorst, 2022).

### 3.4.7. BERTopic

BERTopic is based on the four previously mentioned components, namely, a transformer embedding model, a UMAP dimensionality reduction, HDBSCAN clustering, and cluster tagging using c-TF-IDF. This allows for easily interpretable topics as well as keeping essential words in the topic descriptions (Grootendorst, 2022).

# 4. Method

This section provides a comprehensive overview of the specific methods employed in our project, presenting a detailed account of the specific tools and libraries employed. Each method is presented in chronological order, reflecting the sequence of execution within our project. We also provide insights into our motivations for selecting each tool and highlight their relevance to our project objectives.

## 4.1. Choice of Programming Language

In choosing a programming language, there were a few aspects to consider, such as ease of access for the entire group, availability of libraries, resource intensiveness, and suitability. Python was the only programming language known by the entire group. This fact, along with the abundance of Python-based machine learning libraries, significantly influenced our decision. However, we also had to consider Python's relative computational speed compared to languages such as C++ or C#. After weighing these factors, we decided to use Python because we valued our preexisting knowledge and the available libraries higher than the computational advantages of other languages.

## 4.2. Web Scraping

The first step in the web scraping process was to identify relevant websites and specific search terms from which data was to be extracted. In order to do this, meetings with KAPI were conducted where they provided various social media websites and media channels of interest as well as specific search terms relevant to the topic. Because of time limitations it was essential that we constrain the web scraping process to certain websites. The structure of each website differs from each other significantly, due to this different methods for web scraping had to be implemented for specific websites.

While the web scraping process essentially remains the same, the intricacies involved in scripting a web scraper can vary substantially depending on the platform that is being scraped. Websites like Facebook and Twitter have their own unique structure and representation of data can vary in formats which adds an extra layer of complexity to the scraping process. In addition, such gigantic social media platforms have enough resources to change their structure daily to ensure user experience and security, making the scraping process even more challenging.

Given these complexities and the time constraints we chose not to "reinvent the wheel" by developing our own scraping scripts from scratch. Instead, we turned to GitHub, a platform for software developers to share their code projects publicly. This allowed us to access various web scraping scripts designed for specific platforms, which helped us to accelerate our data extraction process. By benefiting from the collective knowledge and experience from a wide developing community we adapted already existing scripts to meet our specific needs. Even though we leverage some pre-existing scripts, we also acquired a significant amount of practical knowledge and expertise. Our experience facilitated our ability to customise and adapt the pre-existing scripts to meet our specific data collection needs.

Throughout this process, we gained a substantial understanding of various libraries used in web scraping, such as Scrapy, Selenium, BeautifulSoup. We learned the aspect of distinguishing which library would be the most suitable for a given platform, based on the factors such as structure of the website and the nature of the data we need to extract.

Scrapy[2] proved to be a valuable library for building robust and effective spiders, which are used in large scale data extraction. Selenium[3], on the other hand, was more suitable for websites that relied heavily on JavaScript for rendering content. Selenium would imitate the process of manual browsing on a website and automate interactive parts with dynamic content. BeautifulSoup[4] was particularly useful for parsing HTML and XML documents and navigating, searching, and modifying the parse tree, making it a perfect fit for platforms with simpler, less dynamic web contents.

One of the key skills involved in developing a successful scraper is the understanding of XPath. Xpath serves as a locator strategy for navigating through the HTML structure of a given webpage, enabling us to access and manipulate relevant parts of the website. Such

[2] Scrapy. (n.d.). Scrapy | A Fast and Powerful Scraping and Web Crawling Framework. https://scrapy.org/
[3] Selenium. (n.d.). SeleniumHQ Browser Automation. https://www.selenium.dev/
[4] BeautifulSoup. (n.d.). Beautiful Soup: Screen-scraping Parser Tool https://crummy.com/software/BeautifulSoup/

understanding allowed us to effectively navigate and extract desired data from targeted platforms.

Another skill we acquired was the application of regular expressions (REgex), which played a crucial role in our web scraping process. Regex provided a powerful and flexible tool for filtering and processing the scraped data in a relevant manner. It helped us to collect data in almost a surgical manner by picking the parts that are suited for our research and leaving out the irrelevant parts.

However, for some platforms we found that there were no pre-existing scripts available that matched our needs. In these circumstances we took upon ourselves to create tailored scripts. Through understanding the unique structure of each platform we designed a tool for data extraction.

By combining the use of existing web scraping scripts and creating our own when needed, we manage to effectively navigate through the web scraping process on various platforms. Even the pre-existing scripts required some code tweaking and other support-tools to make the scraping possible.

## 4.3. Search Terms

As this project was done on behalf of KAPI, they provided us with a collection of terms they wanted us to investigate. However, depending on the platform that we scraped the complete list could not be utilised. This was because some websites provide more or less sophisticated search functions. Therefore some websites allowed for the combination of the terms, for example, `"Socialstyrelsen" + "kidnappa"`, while others did not.

The usage of different search terms were then motivated by the affordance of the search function for each website. When possible we used, what we call "biased search terms", that is search queries containing the name of a relevant government agency, e.g. `"Socialstyrelsen"` or `"Socialtjänsten" +` a search term from KAPI, such as `"kidnappa"` or `"utredning"`. Biased search terms were used to the largest possible extent. When not possible to use "biased search terms" such as when zero results were found and depending on the internal search engines of the used websites, "unbiased search terms" were used instead. Such search queries simply consisted of the name of a relevant government agency.

## 4.4. Websites

Similar to the list of search terms, KAPI also provided us with an extensive list of websites from which to gather information. However, due to constraints such as time, technical feasibility, and data access, we had to narrow our focus to the following platforms: Twitter, TikTok, Facebook, Reddit, Flashback, and Familjeliv.

### 4.4.1. Twitter

Twitter[5] is a social media platform that enables people to share their opinions and ideas in short messages called "tweets". Users are able to communicate and interact with other users in real-time. Twitter does not allow for longer posts which encourages concise expression. Users often utilise hashtags in their tweets, hashtags categorise their tweets and make them

[5] https://twitter.com/

discoverable when searching for the specific hashtag. The platform serves as a hub for diverse discussions, news and trends.

In order to scrape Twitter we used snscrape[6] which is a scraper for social networking services from GitHub made by the user JustAnotherArchivist. It provides scraping for several different websites such as Facebook, Instagram and Twitter. We found through our testing that for our purpose Twitter was the only website viable through this method. To extract the data we wanted we wrote a script in Python which searches Twitter with combinations of specific search-words together with either `"socialtjänsten"`, `"socialstyrelsen"`, or `"socialtjansten"`.

### 4.4.2. Flashback

Flashback[7] is a Swedish forum with focus on free-speech. The forum has been popular in Sweden since the early 2000s and was founded 1996. The forum uses minimalist design, with little moderation compared to other contemporary websites. Given the small amount of moderation of the forum it follows that it contains more potential disinformation, thus it became an interesting platform for our research purposes.

In order to scrape Flashback we utilised the pre-existing script from GitHub named Flashbackscraper[8]. This script was published by the user christopherkullenberg. Given the complexity and limitation of the targeted platform we had to make modifications in order to complete the scraping process.

The script had the desired functionality to scrape posts from thread URLs. However, the script lacked in terms of usability. Since web scraping is generally an automated tool it has the risk of sending a large amount of requests to the server within a short period of time and many websites implement protection against Distributed Denial of Service (DDoS) attacks by automatically blocking IP addresses that behave in such a way. To navigate around this, we incorporated a proxy into the script to avoid getting blocked. This was done by obtaining a proxy API from ScrapyOps.io services and implementing it into the script. This allowed the script to alternate a different IP address for every request that is sent to the Flashback server.

Although the script was capable of scraping posts from a provided text file containing URLs, it did not have the functionality to extract URLs from desired threads. To address this, we developed a script that was able to scrape all threads based on predefined search terms and export them into a text file.

With the text file ready, we were able to scrape all posts from targeted threads. However, the scraped data contained a significant amount of noise or irrelevant information along with the posts. Additionally, the script exported scraped data to a SQLite3 file which was not suitable for our analysis.

To resolve such issues, we had to write another script that was able to process the SQLite file. The script's purpose was to extract only the post messages and their corresponding dates, while ignoring the rest. Finally the script exported the targeted data into a CSV file providing us with the desired format for future analysis.

[6] https://github.com/JustAnotherArchivist/snscrape
[7] https://www.flashback.org/
[8] https://github.com/christopherkullenberg/flashbackscraper

### 4.4.3. Reddit

Reddit[9] is the world's largest discussion platform. While Reddit's user base is largely located in the USA, there are enough Swedish users to make it a viable source of data for our research.

To scrape Reddit we used the scraper named URS[10], which was found on GitHub and published by user JosephLai241. The scraper operates via PRAW, the official Python Reddit API wrapper, which utilises an API provided by Reddit to make scraping possible. By registering an account on Reddit's developers page we acquired an API and integrated it into the script.

The scraper came with several functionalities, however, we were primarily interested in its ability to scrape data using predefined search terms. One of the drawbacks of the Reddit scraper was that it could only use such a feature in conjunction with specific subforums, known as subreddits, within Reddit. This meant that we had to manually identify subreddits to scrape using predefined search terms. By doing so we identified several recurring subreddits which were later targeted for data scraping.

While the Reddit scraper was able to scrape URLs of threads within the chosen subreddits, it also extracted a large amount of irrelevant information. To deal with the issue, we developed a script to isolate the URLs and export them into a text file.

Another functionality of the Reddit scraper was to extract all posts within a single thread. This could be accessed by providing a URL for a given thread. Given our extensive list of URLs, we had to develop a script that would automate the process by feeding the list of URLs in order to scrape all data from every URL. This allowed us to scrape a significant amount of posts from each targeted thread. However, similar to the URL scraping process, the post scraping also yielded a lot of noisy data. Consequently, we wrote another script that would isolate only the posting date and the post message from each scraped post. Lastly, formatted data was stored into a CSV file for further use.

### 4.4.4. Familjeliv

Familjeliv[11] is a highly frequented online platform in Sweden that garners more than a million daily visitors. This makes the platform a well established internet forum in Sweden, with diverse discussion topics, politics being amongst the top. While Flashback tends to capture a younger crowd, Familjeliv provides access to a much more diverse age group which makes it a viable source of data.

Because no pre-existing script was available on GitHub we had to construct a custom web scraping script from scratch. The web scraping process was divided into two steps for the sake of simplicity. Familjeliv organises the posts throughout several pages within a particular thread. Our aim was to extract all posts within a relevant thread.

The first step involved creating a script utilising the Selenium library in Python. The script purpose was to navigate through Familjeliv forum section and extract URLs of all the threads that corresponded to our predefined search terms. All URLs were compiled and stored into a

[9] https://www.reddit.com/
[10] https://github.com/JosephLai241/URS
[11] https://www.familjeliv.se/

single text file. Some search terms could yield already existing URLs in the text file. To avoid this, the script had an additional feature built into skip duplicates.

In the second step, we developed a web crawler based on the Scrapy library in Python. The webcrawler had the ability to iterate through every URL inside a text file and extract messages from each post along with their corresponding posting dates. The extracted data was exported into a single CSV file, organised by two columns “date” and “messages".

This two-step approach allowed us to handle the complexity of Familjeliv’s structure and effectively scrape the required data.

### 4.4.5. Facebook

Amongst the websites KAPI requested us to scrape were Facebook[12], which currently is one of the largest social media platforms globally. It provides a platform for maintaining relationships, sharing life updates and staying connected with others. Facebook offers functions beyond social networking, such as community building. Here, users can create and join groups based on shared interests, facilitating discussions on various topics including experiences and perspectives related to organisations like Socialstyrelsen.

To collect data from Facebook, the Python library Selenium[13] was used. A script was developed that searched for posts mentioning the relevant government entities, i.e “Socialstyrelsen, Socialtjänsten, and other aliases for these. From these posts the script extracted information from the post itself and its comments by using xpaths. Furthermore, several groups of interest were found. These groups contained a great amount of discourse that we deemed as pertaining to the LVU-campaign. Thus a second script was developed that collected data from such groups by employing the same method as the previously mentioned one. Both of the developed scripts output the data into a CSV file.

### 4.4.6. TikTok

TikTok[14] was one of the websites requested for scraping by KAPI. TikTok is a massive social media platform, boasting hundreds of millions users worldwide. The platform enables users to create and share short videos with other users. It provides a diverse range of content categories such as dancing, comedy sketches, political content and more. TikTok also allows users to comment on each others videos.

TikTok is primarily a video-sharing social media platform. Although we are not interested in the videos themselves for this project, the comments on these videos are valuable. To collect these comments, a JavaScript script, created by the user cubernetes[15] and obtained from Github, was executed using the Chrome DevTools console (used to run code on websites). Scraping was performed using the following steps. Initially, we navigated to videos of interest by accessing TikTok and using its search function to look for “socialtjänsten”. This yielded numerous relevant results suitable for our analysis. Subsequently, we opened each video and accessed Chrome DevTools, into which the JavaScript was pasted. The script automatically saved and copied each comment to the clipboard, enabling us to paste the comments directly into a CSV file.

---

[12] https://www.facebook.com/
[13] Selenium. (n.d.). SeleniumHQ Browser Automation. https://www.selenium.dev/
[14] https://www.tiktok.com/
[15] https://github.com/cubernetes/TikTokCommentScraper

## 4.5. Pre-Processing

Automated analysis techniques like topic modelling and semantic analysis require consistently formatted data. Writing a Python script that dynamically adapts to diverse date formats or column content is considerably more challenging than first standardising the data before analysis. This standardisation enhances reliability. Beyond just ensuring data consistency, this step also involves detecting and correcting errors, inconsistencies, and outliers.

### 4.5.1. Initial Examination

We initially examined the data by manually going through our scraped data and visually inspecting it. The data was stored in different CSV-files depending on which forum and script they were gathered using. So depending on what approach the data was scraped with and where it was scraped from, for example, Twitter with search terms, Twitter based on profiles, Flashback without search terms etc. During this visual inspection there were four main things which were initially obvious. Firstly, the date formats were incredibly varied, ranging from DDMMYY, YYYYDDMM, MMDDYY, to "older than YY-DD-MM". Secondly, certain elements related to the internal formatting of websites, such as hashtags, @s, /r/, /u/ and CITAT: were quite prevalent. These needed to be removed to avoid potential confusion or misinterpretation during the analysis, given their specific context-dependent meanings Thirdly, we had several URLs in our dataset, which would serve only to confuse our topics. Fourthly, we had a significant amount of duplicates within our database, to remove these we used the Python library Pandas[16], as we needed to work on a file containing ~ 400 000 social media posts and comments.

### 4.5.2. Data Screening

Our criteria for data were quite simple. Firstly, all files are CSV files. These CSV-files all have the header date, lemma, text wherein date is formatted YYYYMMDD with no - or / such as 20010926, according to the ISO 8601 basic format[17]. Lemma holds the base form of each word in a post, while text carries the original post content, having removed URLs, website-specific formatting, emojis, and words containing non-ASCII characters, with the exception of Å, Ä, and Ö. These criteria were decided upon since we had to change from several datetime formats, and therefore found it easiest to remove all unnecessary formatting.

### 4.5.3. Handling Missing Data

Handling missing data was a relatively straightforward process given the volume of our dataset, which consisted of over 400,000 social media posts and comments. We decided to discard any posts that lacked either a date or textual content. Since our data only consisted of these two key elements, the number of such incomplete posts was quite low. This simplified our decision-making process compared to if there had been more incomplete data. As we aimed to focus specifically on potential misinformation, which was likely more prevalent in our dataset during the months of the LVU-campaign, a significant filtering method we adopted was to examine data strictly from that period. However, we used the entire dataset as a baseline to discern how discourse evolved during the peak of misinformation spread. Given

[16] https://pandas.pydata.org/
[17] https://www.iso.org/iso-8601-date-and-time-format.html

this, the small portion of data we omitted was a justified trade-off to ensure the effectiveness of our filtering method.

### 4.5.4. Anomalies & Standardisation

After this initial visual examination we discovered less common quirks while debugging the code and trying to use the partly cleaned data. These quirks varied from platform to platform, such as some posts being completely empty, this was a result of a user posting only spaces and certain posts having been removed only showing [deleted] or [removed] in our data. Since emojis can be used by two different individuals to convey significantly different meanings, we decided they also had to be removed. All of the actual cleaning was done using different Python scripts, primarily through the use of different regex-expressions. For example `"Citat:\s*Ursprungligen\s+postat\s+av"` which matches "Citat: Ursprungligen postat av". We then use the `re.sub()` function call to replace this text with an empty string. This was done using Python's `re` module, which is used to handle regex. Another used regex was `"\/[ur]\s\/"` which corresponds to both "/u/" and /r/ using regexes `[]` operator. A third regex used was `"@\S+\s?"` which matches a @ and the following word, this was specifically used to remove mentions on Twitter. This is because we limited ourselves to only working on Swedish, and this conveniently excludes non-latin based characters. Duplicates were also removed using the Pandas library, this was because we scraped the same websites using different approaches which brought content which sometimes overlapped. Say we first scraped a certain hashtag, and then indexed the account which posted with said hashtag, we would end up scraping at least one post which overlapped. Thankfully the Pandas library is incredibly well suited to taking a large dataset and removing duplicates.

### 4.5.5. Quality Control

Given that our analysis requires data to be formatted in a very specific and consistent manner, our primary quality control measure involved running verification programs to ensure each date in our CSV was correctly formatted. Alongside this, we also conducted a visual inspection. While this method didn't catch any additional issues, it served as an extra layer of quality assurance. The inspection was carried out by scanning a few thousand consecutive lines in the CSV, then skipping roughly an eighth of the file, and repeating the process. This method ensured we had inspected data from each individual CSV that contributed to the larger file. We did not inspect individual files this way due to an oversight and the anticipated time inefficiency of such a method.

### 4.5.6. Ethical Considerations

We initially scraped usernames while we were gathering data, but due to ethical considerations in conjunction with time constraints we decided to remove this data. However this does not mean our dataset is free from names and usernames, as several of our textbodies include names, usually of politicians or other public figures. This does not mean we have specifically gathered online statements from such persons, instead they are themselves topics of discussion due to their public stances and opinions. Therefore we feel removing these discussions is not necessary from an ethical perspective. As these are relatively few in number they do not impact our topics to the extent that they would be worth removing, compared to other things. Also from an anonymization perspective we do not include whichever person made a statement only someone else's statement regarding these figures.

## 4.6. Analysis

This section will present our method for analysis on the collected data.

### 4.6.1. Sentiment Analysis Using VADER

To perform sentiment analysis on our dataset, we employed the VADER tool. Firstly, the text data was fed into VADER, which then processed the text. VADER scanned each document for words that existed in its predefined sentiment lexicon, and assigned a sentiment polarity score to each word based on this lexicon. The polarity score was then adjusted based on the context, which was derived from the surrounding words, and the presence of intensifiers or diminishers.

In the case of emoticons, slang, or acronyms, VADER used its specially designed scoring and mapping system to assign sentiment scores. This ensured a more accurate representation of the sentiment, given the informal nature of social media language. The final output for each document was a composite score, calculated by summing the individual word scores and adjusting for sentiment intensification and grammatical rules. This composite score represented the overall sentiment of the document.

### 4.6.2. Topic Modelling Using LDA And BERTopic

Following the sentiment analysis, the cleaned and pre-processed data was passed through our topic modelling pipeline, which consisted of LDA and BERTopic.

The LDA model was applied first. Each document was represented as a mixture of various topics and the topics were represented as a distribution of words. The LDA model then generated topics by iterating through each document and assigning each word to a topic. This assignment was based on the probability of the word belonging to a topic and the probability of the document generating a particular topic.

Despite its effectiveness, LDA tends to neglect semantic relations between words. To counter this, we employed BERTopic. BERTopic starts by converting the text data into dense vector representations using the paraphrase-multilingual-mpnet-base-v2 transformer model. These dense vector representations capture the semantic meaning of the text and the context in which words are used.

Next, BERTopic reduces the dimensionality of these embeddings using UMAP for easier processing and better performance. The dimension-reduced data is then clustered using the HDBSCAN algorithm, effectively grouping similar data points, which, in this case, represent documents. Finally, for each cluster, c-TF-IDF is used to extract the most representative words, which are then tagged to the respective clusters. These tagged words form the description of the topics.

By cross-examining the results of the sentiment analysis and topic modelling, we were able to associate the prevalent sentiments with the discovered topics, providing a comprehensive understanding of our dataset.

## 4.7. Qualitative Analysis

The qualitative analysis of the data was conducted in order to gain insights into the content and characteristics of the dataset. The analysis consisted of several steps, including the identification of topics of interest, the identification of periods of interest within these topics, the classification of posts as potential disinformation, and a content analysis of the posts classified as potential disinformation.

### 4.7.1. Topics Of Interest

To identify topics of interest the topic labels provided by BERTopic, which automatically assigns labels to clusters of related posts based on their content, were inspected to identify topics specifically related to the LVU-campaign. The terms related to the campaign were provided by our employer, KAPI, and gathered from the report on the LVU-campaign produced by Ranstorp & Ahlerup (2023).

Topics deemed as interesting for further analysis were subsequently divided into separate datasets, facilitating subsequent analysis.

### 4.7.2. Periods Of Interest

Identifying periods of interest was done in order to provide a better overview and categorisation of each identified topic for KAPI.

To identify periods of interest the datasets for topics of interest were iterated over, where significant shifts in the overall discourse in the user-generated posts constituted the beginning of a new period of interest. Additionally, the temporal aspects of the data were analysed, such as fluctuations in the number of posts or changes in sentiment, to further aid in the identification of periods of increase . Typically, an increase in the number of posts or a notable shift in sentiment indicated the beginning of a new period of interest.

### 4.7.3. Classifying Potential Disinformation

After this was done, each document containing potential disinformation was annotated. A post was deemed as potential disinformation if the narrative in the post aligned with common narratives of the LVU-campaign and expressed an overly negative sentiment towards Socialstyrelsen. These common narratives, used for classifying a post, were gathered from section 4 in the report on the LVU-campaign authored by Ranstorp & Ahlerup (2023).

### 4.7.4. Content Analysis

After annotating the posts containing potential disinformation, a content analysis was conducted to extract the specific details regarding what was being said about Socialstyrelsen, how it was being conveyed, and when it occurred. This analysis delved into the textual content of the posts, examining the language used, the tone employed, and the timing of the discussions. The purpose behind analysing the content was to be able to extract and present the content of the potential disinformation being spread and present these findings to our employer.

# 5. Results

This section will present our findings obtained from the analysis of the collected data. Starting with a general overview of the data and followed by delving into the specific topics identified, providing detailed analyses, representative excerpts, relevant visualisations, and identified periods of interest that were identified during the qualitative analysis, to facilitate a comprehensive understanding of the discourse.

## 5.1. General Data Overview

Before delving into topic-specific analyses, a general data overview is presented of the data collected for this study. The dataset used for analysis, that is, posts from 2021-01-01 and forward, comprises a total of 87,198 posts. In total, BERTopic identified 619 topics, the distribution of topics is represented in Figure 2. The reason for using data from 2021 and onward is because this is the period of interest from our stakeholders. The negative topic, that is, data entries that did not fit into any topic, consists of 48,988 entries. The negative topic comprises 56.2% of the collected data from 2021 and forwards. Out of 619 topics, three were identified as relevant to the LVU-campaign, these topics were selected on the basis of their top words and content. The largest topic cluster (excluding the negative topic) contains 1595 entries while the smallest contains 10 entries. The average number of data entries for each topic is 61.83 and the median is 24.5 (Std. dev = 149.26) and the average sentiment for all the data is -0.08.

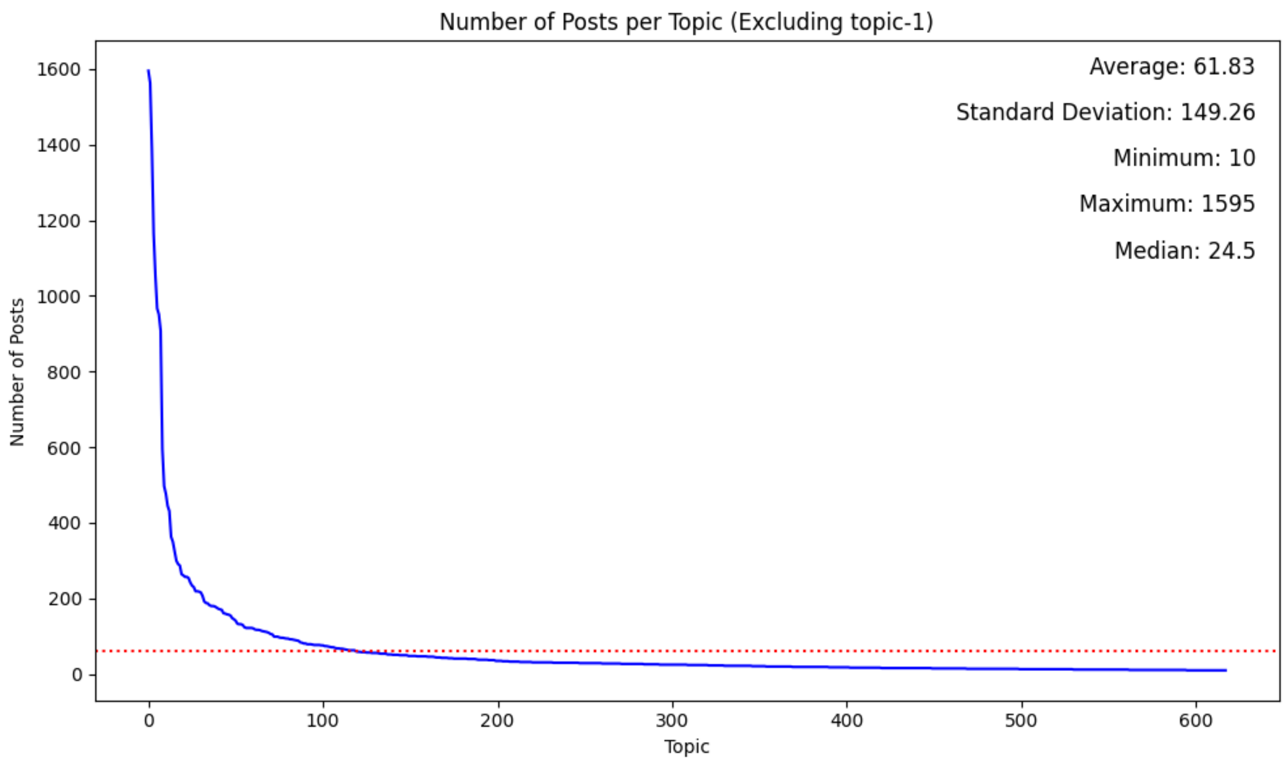


*Fig 2. Visualisation of topic distribution*

Furthermore, the dataset is characterised by diverse content, encompassing a range of perspectives and opinions. The posts encompass discussions, reactions, user-generated posts, and instances of misinformation, which necessitates a careful examination to discern the nuances and identify the prevalence of potential disinformation. The analysis will shed light on the representation of such posts within the dataset.

## 5.2. Topic 11

Topic 11 referred to as muslimsk_kidnappa_muslim_islamist by BERTopic contains 445 posts and makes up 1.2% of all data, excluding the negative topic. To summarise, the topic mainly consists of user-generated content discussing, reacting and/or sharing news articles related to the LVU-campaign. The average number of posts per month for this topic was 18.5 with February 2022 containing the highest number of posts, as seen in Figure 3.

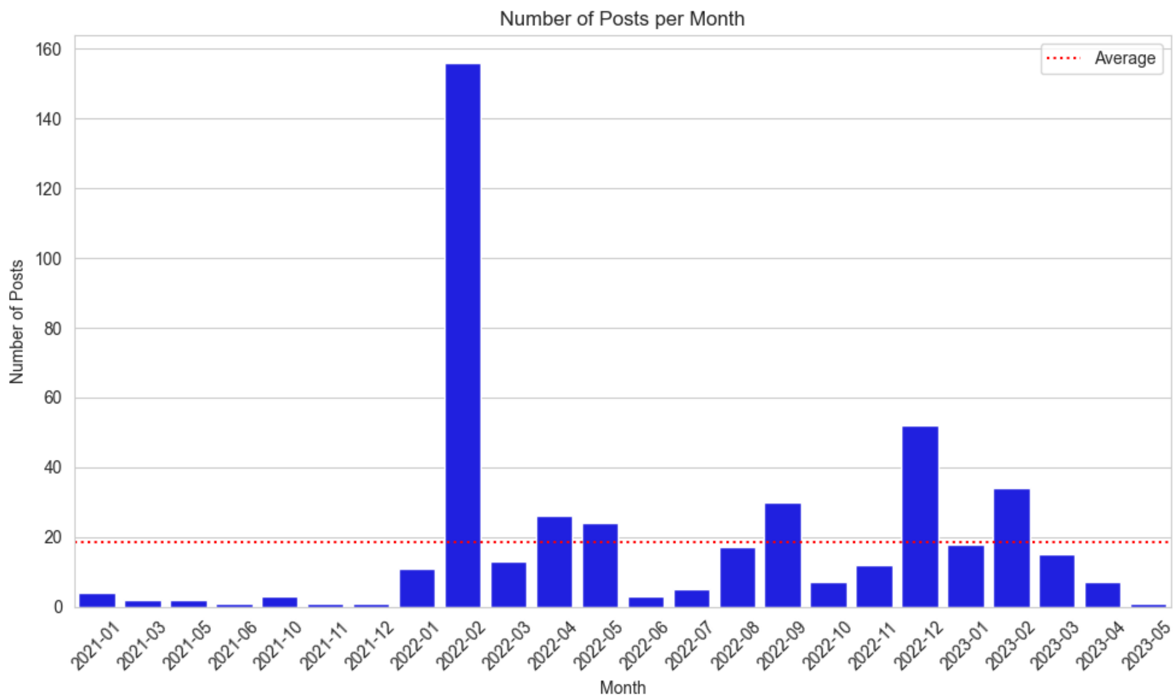


*Fig 3. Distribution of posts for topic 11*

### 5.2.1. Representative Topic Posts

To gain a better understanding of the topic, three carefully selected data points are presented below. Accompanying each post is the date it was posted, the probability of it belonging to the topic, if the post is representative of the topic and its sentiment score. The rolling sentiment average sentient score can be seen in Figure 4.

| **Date** | **Probability** | **Representative** | **Sentiment score** |
|---|---|---|---|
| 2022-02-06 | 1.0 | False | -0.36 |

Enligt en av de mest drivande för denna kampanj för smutskastning av socialtjänsten att "" omhändertagandet av muslimska barn är välplanerad och systematisk mot muslimer för att minimera islams tillväxt och muslimska traditioner inte förs vidare ""

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-01-27 | 1.0 | False | -0.73 |

"En uppfattning som sprider sig .
Socialtjänsten kidnappar muslimska barn för att sekularisera dem som en del i kriget mot islam . Välkända namn inom den radikalislamistiska miljön deltar och eldar på misstron mot samhället och tjänstemän .

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-01-27 | 1.0 | False | -0.72 |

""" Uppfattningen att Sverige kidnappar barn är inte bara utbredd bland de mest extrema islamisterna .
I samhället sprids lögnerna bland vanliga medborgare som tror att socialtjänsten kidnappar barn , tvingar dem att äta fläsk och ha sex . ""

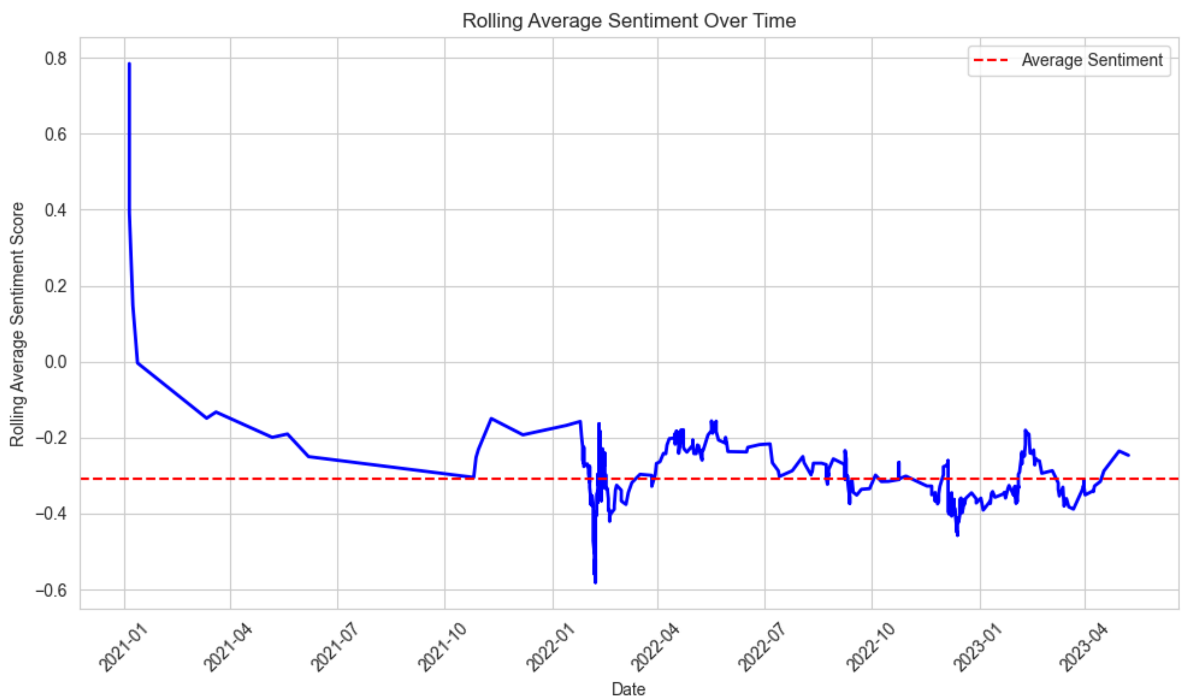


*Fig 4. Rolling average sentiment over time for topic 11*

## 5.2.2. Period 2021-01-01 To 2022-01-27

During the period of 2021-01-01 to 2022-01-27 there is a relatively low amount of posts in the dataset. The discourse during this period mainly consists of anti-immigrant, anti-muslim and xenophobic views. Contained within this period is, however, one post that we deemed as potential disinformation. It was posted 2022-02-25 and reads:

*"socialstyrelsen There is a Palestinian family saying that the socialstyrelsen has kidnapped their 5 children , ironically , the father received pictures of his 13 years old daughter being drunk ! Any explanation ?"*

A possible significant event during this period, that is mentioned 2022-01-13 is the case of two parents charged with the crime of aiding and abetting genital mutilation accuses Socialstyrelsen of targeting muslim families. The post is shown below along with a reference to the relevant article.

*Ur ________ text : "" Föräldrarna talar om hur socialtjänsten aktivt vill komma åt och förstöra för muslimska familjer , att könsstympning är seder och traditioner . En vän säger att han tränat sin dotter i vad hon ska säga när känsliga ämnen kommer upp i skolan .*

Around the end of this period, the first mention of the LVU campaign against Socialtjänsten is mentioned, specifically 2022-01-26. However, the reason for the significant increase in the amount of posts, as seen in Figure 3 is likely attributed to the release of a news article. The article, released by an independent news agency 2022-01-27, coincides with an increase in posts. Furthermore, as other news sources also began reporting on the campaign the activity within the topic further increased.

### 5.2.3. Period 2022-01-27 To 2022-09-08

The next period that was identified contains mainly user-generated content related to the LVU-campaign. Sentiment, in the beginning of this period, is somewhat neutral as posts mainly quote articles. However as a discussion about the campaign emerges, the sentiment quickly decreases, as seen in Figure 4. This steep decrease, we attribute to users reacting to the news about the campaign.

As a whole the period contains 18 posts that we could classify as potential misinformation, whereof two similar yet different narratives were identified. Contained within the first narrative is the accusation of Socialstyrelsens rampant kidnapping of children with muslim background. The second narrative, on the other hand, is that the legal framework of LVU and courts is questionable and has been used not only for the kidnapping of muslim children but children with Swedish backgrounds as well.

As the frequency of user-generated content decreases during the latter part of the period the disinformation narratives continue to be present within the data. During May the narrative present in this topic of Socialstyrelsen kidnapping muslim children expands to include the act of forcibly converting them to Christianity. Also as the election nears, the amount of mentions of political entities in relation to the campaign increases.

### 5.2.4. Period 2022-09-08 To 2022-10-24

During this period, the discussion is mainly about the campaign in relation to the Swedish election, the period contains 36 posts. Interesting developments related to the existing LVU-campaign in relation to mainly two political parties occur during this period. Mainly, the involvement of the political party Nyans in spreading misinformation is mentioned. Nyans, a minor Swedish political party self described as: “an inclusive and pragmatic party with the goal of solving commonly encountered problems for people in minority groups” (Partiet

Nyans, 2020) is accused by users within the data set of echoing the same narratives as those from the LVU-campaign.

Another interesting event mentioned in the period by users is that a politician from Vänsterpartiet, a left-wing party in Sweden, gave a speech within which the same narratives were shared. It is important to note that not all the posts express support for the campaign, as the discussion revolves more around the spreading of information and its various perspectives.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-09-08 | 0.77 | False | -0.4 |

"_____ Men V som sprider konspirationsteorier om att socialtjänsten tvångsomhändertar muslimers barn har du inga problem att samarbeta med . "

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-09-08 | 0.77 | False | -0.03 |

"vansterpartiet ______ Eftersom jag ska rösta på söndag och vill göra ett genomtänkt val skulle jag uppskatta en kommentar om det är Vänsterpartiets linje att Liberalerna vill att socialtjänsten ska kidnappa muslimska barn ? Det är er justitieministerkandidats uttalande ”

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-09-08 | 0.47 | False | -0.08 |

"Hon bekräftar partiet Nyans som säger att muslimer systematiskt får sina barn omhändertagna av svenska myndigheter . Hon spär på den misstro som finns mot socialtjänsten och hon försvårar socialtjänstens arbete . Varför _____ varför ? & gt ; svpol “

### 5.2.5. Period 2022-10-24 To 2023-05-08

After a short intermission during the election period, the discourse shifted back to mainly pertaining to the LVU-campaign. Furthermore, during this period, a development in the disinformation narrative emerges, namely that muslims are encouraged to leave Sweden. Our explanation for the increase in posts in December 2022 is attributed to this development. The first mention of Muslims being encouraged to leave Sweden occurs 2022-11-28 in the post on the next page.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-11-28 | 0.46 | False | 0.14 |

" Duties of Muslims in Sweden : - Not to remain silent about their rights nor leave their children , but rather they must leave those countries . "" Socialtjänsten

Later it seems that a second wave in the disinformation campaign begins around February 2023. Worth mentioning is that the scraped data for Twitter drops off around the end of May, as the scraping tool used stopped working. Therefore, as much of the disinformation-campaign and its related discussion took place on Twitter this could explain the drop off in posts around the latter part of this period.

### 5.2.6. Content Analysis

Present in the data set for topic 11, distinct narratives relating to the disinformation campaign were identified. The first narrative is centred around the kidnapping of children, specifically targeting children of muslim background. In relation to this, allegations of forced conversion to Christianity are mentioned through placement in christian homes. Criticism of Socialstjänsten and its handling of child welfare cases are prevalent throughout the topic and references to conspiracy theories involving government officials and child trafficking. The discussion around how Socialstjänsten handles child welfare implies that the agency is racist and islamophobic. These allegations also extend to Sweden as a whole, where it is expressed that Sweden is a racist and islamophobic country. This narrative later develops into that muslims are encouraged to leave Sweden.

Another narrative identified can be summarised as Socialstyrelsen and Swedish courts operating on a legally questionable basis. This narrative emerges at the same time as the first narratives in line with the LVU-campaign. The narrative was classified as potential disinformation as the sentiment in many posts seem to agree with the first narrative. Another interesting claim within this narrative is that the kidnapping of children isn't anything new in Sweden and has occurred with other minorities as well during Sweden's history. This acts as a basis for many posters to agree with the first narrative. These claims are generally more factually supported (e.g see footnotes [18] & [19] ), as mentioned however, because many posters seem to agree with the LVU-campaign we deemed this narrative interesting for our employers further analysis.

## 5.3. Topic 52

Topic 52 referred to as kidnappa_kidnapping_barn_kidnappare by BERTopic consist of 113 posts and make up 0.002% of all data, excluding the negative topic. A detailed examination of the data reveals that the discussions mainly revolve around alleged instances of child kidnapping by Socialtjänsten. In total the number of posts that were classified as potential disinformation is 41, making up 31% of the entire topic data. The average number of posts per month for the given topic is 8.25 with sudden spikes in Februari, November, December 2022

[18] (”Hur Staten Hanterat Samer är Fruktansvärt Dåligt”, 2020)
[19] (Historiska Paralleller Till Hur Romer Behandlas Idag - | Forskning.se, 2015)

and January 2023, presented in Figure 5. The rolling average sentiment over time for topic 52 is represented in Figure 6.

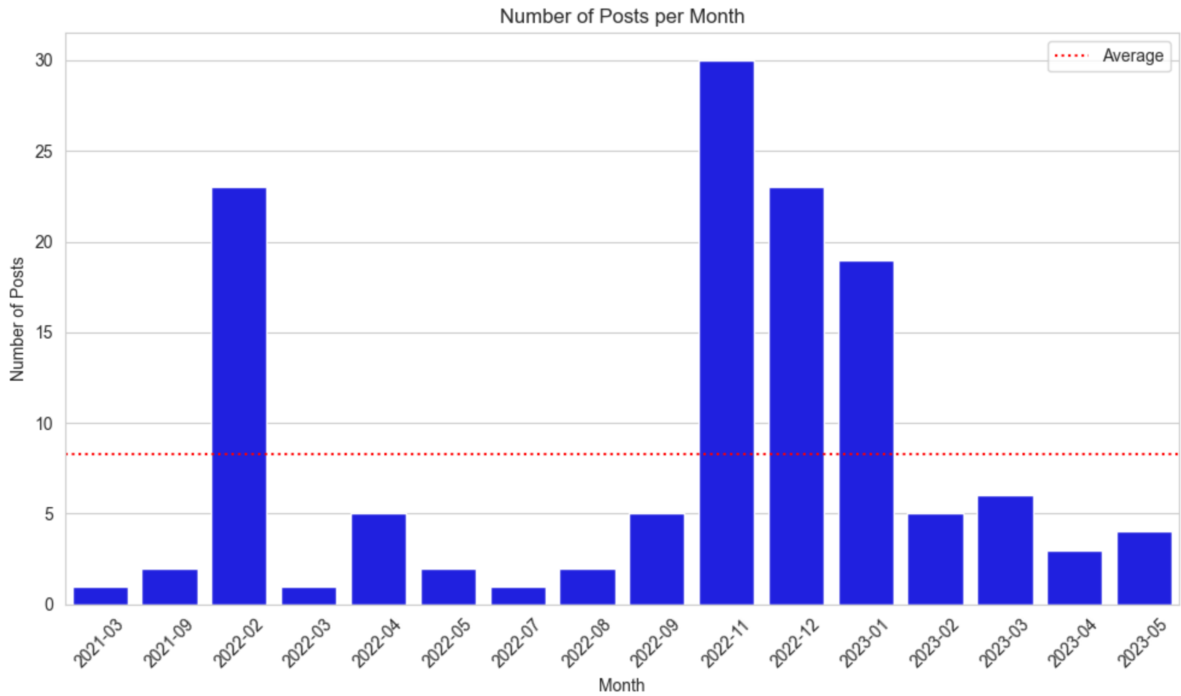


*Fig 5. Distribution of posts for topic 52*

### 5.3.1. Representative Topic Posts

To gain a deeper understanding of topic 52, the following posts have been meticulously selected as representative samples. BERTopic assigned these posts a 100% probability of belonging to this topic. However, despite the high probability score, none of the posts were marked as representative by BERTopic. This occurrence was not unique to these posts, as was the case for all the posts within topic 52.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-14 | 1.0 | False | -0.5 |

"Expressen Döda dem själv deras barn så socialtjänsten kan inte kidnappa dem

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-11-19 | 1.0 | False | -05.6 |

"Tror du att hon kommer till ett nu ? gjorde inget annat att kidnappa detta barn och krossa barnets .

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-11-25 | 1.0 | False | -0.1 |

"Sen hjälper polisen också socialtjänsten att kidnappa barn från familjer

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2023-01-28 | 1.0 | False | -0.12 |

"Men om det är socialtjänsten som är i färd med att kidnappa dina barn ? Går du till anfall ? Det får jag hoppas att du gör

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2023-01-12 | 1.0 | False | 0.49 |

"Hur kan barnen bli omhändertagna och sedan därefter bli förda ur landet ? Kidnappades de från ett hem eller ?

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-07 | 1.0 | False | 0.77 |

"Hur många barn har socialtjänsten kidnappat egentligen ? Ska bli intressant att se hur stor uppslutningen blir

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-12-06 | 1.0 | False | -0.46 |

" Samtidigt som socialtjänsten kidnappar deras barn

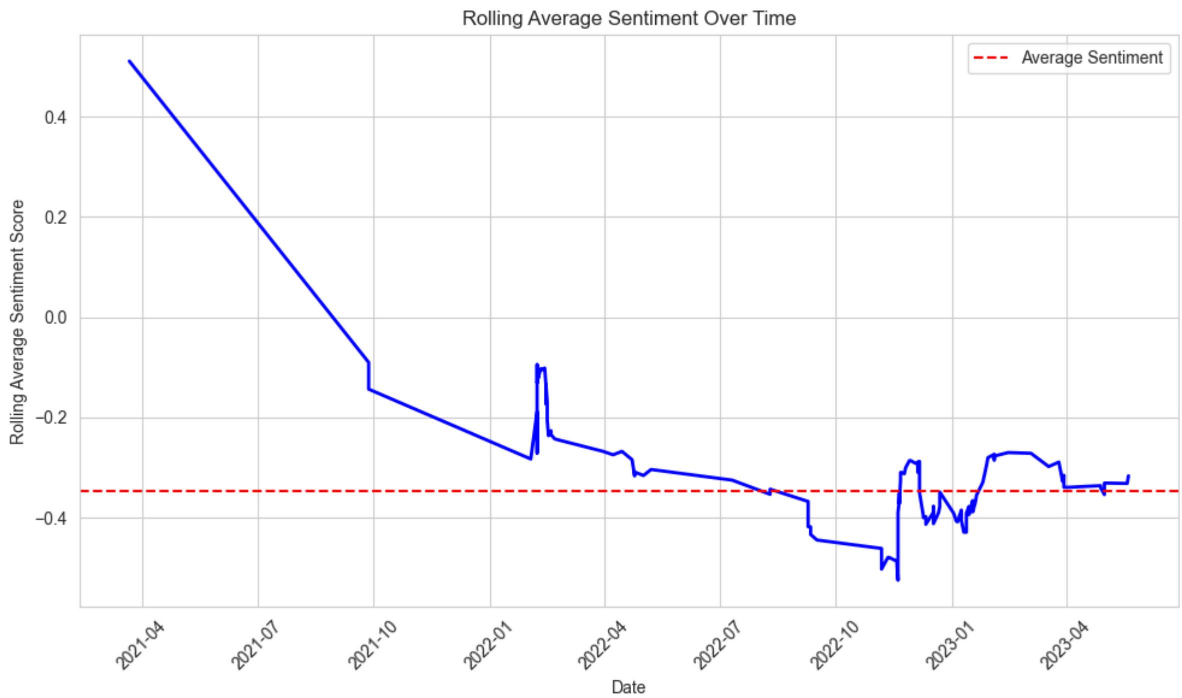


Fig 6. *Rolling average sentiment over time for topic 52*

### 5.3.2. Period 2021-03-22 To 2022-02-21

The posts spanned from March 2021 to February 2022 consisting of 26 posts in total while 23 of them are from February 2022, indicating a sudden spark in discussion. This period provides a look into the sentiment regarding operations of Socialtjänsten, particularly in relation to child kidnappings.

A potential narrative identified in the posts suggesting an accusatory stance towards Socialtjänsten. These posts suggest that the Socialtjänsten is acting based on misinformation or unjustly removing children from their parents. Such views are particularly visible in the following data.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-02 | 0.42 | False | -0.7 |

"Stöttar utsatta föräldrar som fått sina barn kidnappade av socialtjänsten baserat på lögner .

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-11 | 0.56 | False | -0.13 |

"För att bevisa att socialtjänsten inte kidnappar barn , ska alltså en "" nationell enhet "" kidnappa barn ?
Probability: 0.56 Representative: False Sentiment score: -0.13"

Additionally, some posts convey intense emotions and frustration, painting the Socialtjänsten in a highly unfavourable light. This can be observed in the posts below. The sentiment score remained neutral for the posts written in English since the VADER was based on Swedish lexicon.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-05 | 0.28 | False | 0.0 |

"socialstyrelsen You are devil ! ! ! Stopkidnappingourchildren

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-07 | 0.47 | False | 0.0 |

"Stop kidnapping our Kids socialstyrelsen

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-14 | 1.0 | False | -0.5 |

"Expressen Döda dem själv deras barn så socialtjänsten kan inte kidnappa dem

This period also included some sceptical views towards the accusative stance against Socialtjänsten. Such posts called into question the accusatory claims made against Socialtjänsten and suggested the possibility of misinformation.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-28 | 0.64 | False | -0.13 |

"Vilka barn kidnappar socialtjänsten och har du någon källa på vem som sprider dessa rykten ?

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-14 | 0.13 | False | 0.59 |

"Precis så är det .. fattar inte hur folk ens kan tro att barn kidnappas av socialtjänsten .. tror man på troll oxå ? Ser inte mycket i debatten av att det ska handla om barnets bästa .. och enbart det .. oberoende av vilken folkgrupp man tillhör .

This period is also accompanied by posts that stand up for Socialtjänsten arguing that child removal procedures were only carried out on legal basis and for the child's best interest. Such as in the following post.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-15 | 0.19 | False | -0.94 |

"- Socialtjänsten anklagas för att kidnappa barn . Men att omhänderta dem från föräldrar som gör dem illa är rätt . Att politikerna inte lyckats förmedla vad som gäller här med pondus är en skam .

Another post criticised the notion of malicious child removal carried out based on racial grounds.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-13 | 0.12 | False | -0.05 |

"Att folk som borde vara sansade och intelligenta på allvar tror att socialtjänsten skulle okynneskidnappa barn utifrån hudfärg är så bisarrt . Särskilt när i princip alla landets sockontor larmar om att de inte hinner med att utreda de orosanmälningar som kommer in .

Conclusively, this period indicates a divide in public opinions and negative tone of discussion concerning Socialtjänsten.

### 5.3.3. Period 2022-03-31 To 2022-09-16

Following period takes place after the sudden spark of discussion regarding child kidnapping carried out by Socialtjänsten. This period stretches over seven months and consists of 16 posts in total. While the discussion is sparse, there is some support that people refer in past tense to what was happening in February 2022. Following posts were found with such a notion.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-04-08 | 0.35 | False | -0.46 |

"Detta säger inte allt . Knappt nåt . Det finns människor som tror att socialtjänsten kidnappar och konverterar barn . Säger det allt om socialtjänsten ?

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-05-02 | 0.18 | False | -0.46 |

"Och sen när barnen blir placerade så blir det ännu fler som kommer att stå och gapa utanför socialtjänsten om att dom kidnappar deras barn , våldtar och säljer dom . Heeelt utan anledning .

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-08-10 | 0.16 | False | -0.85 |

"Galet mycket påhopp från vänster om att barn ska tvångsomhändertas och att barn ofta inte kan prata som 2-åringar . Skrämmande att se vilken argumentation man använder . Påminner mycket om kampanjer från Mellanöstern kring att socialtjänsten kidnappar barn .

This period also contains a distinctive subset of posts that employ sarcasm or satire in address to the narrative of the Socialtjänsten kidnapping children. Such a humorous and dismissive approach is suggestively aimed at criticising those who support the narrative and highlight the absurdity of those who support these claims.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-04-23 | 0.69 | False | -0.76 |

"Åh herregud , nu blir det kravaller med ""Socialtjänsten kidnappar våra barn ! ""

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-04-25 | 0.96 | False | -0.46 |

"Låter som att treåringen har ett extremt behov av att bli kidnappad av socialtjänsten .

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-04-25 | 0.74 | False | -0.46 |

"Jaha , va trångt det måste vara för jag hörde att hon hade en bur fylld med kidnappade barn från socialtjänsten där under .

### 5.3.4. Period 2022-11-06 To 2023-05-20

Following period begins with another spark in discussions by a total of 72 posts in a course of the first three months. While no distinctive narrative can be observed in this period, the data suggests that people continue with a divided stance on the kidnapping narrative. However, the sentiment is tilted toward a more negative level of -0.34 in average sentiment score when compared to February 2022 being at -0.28 suggesting stronger division between opposite views and harsher tone of discourse such as in the following posts.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-12-06 | 0.12 | False | -0.91 |

"Nu kallas Socialtjänsten kidnappare . Social sekr hotas . Barn lider ! I debatten talas om desinformation . Men nej ! Det är en reell kulturkrock ? De
har en annan uppfattning än våra lagar uttrycker och accepterar inte vår syn inte bara här utan i många andra frågor också . Inse detta !

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-12-11 | 0.39 | False | -0.73 |

"Ytterligare ett bevis att socialtjänsten har en agenda . Dom skiter i barnen och föräldrarna . Dom är bara ute efter att kidnappa och söndra

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-12-03 | 0.3 | False | -0.68 |
| "Därför att rättsväsendet blivit politiskt ! Se bara på alla grundlösa anklagelser mot föräldrar där socialtjänsten kidnappar deras barn | | | |

Level of discussions decreased after January 2023 down to 18 posts for the remaining period while the sentiment further decreased to -0.42. Some nuances are introduced, like the attempt to redefine the term "kidnapping" providing an alternative perspective on the discourse. This can be seen in the following posts.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2023-02-03 | 0.44 | False | -0.69 |
| "Ordet kidnappa syftar väl på att det rör sig om ett olovligt bortförande eller frihetsberövande . Sen bör man nog inte tala så mycket om "" hårda konsekvenser "" eller liknande . Om man inte kan sköta vårdnaden av sitt barn så kan det leda till att barnet tvångsvårdas ja , men det görs för att skydda barnet och det är ingen bestraffning . | | | |

Conclusively, the public opinion seems polarised on the entire narrative. However, the conversation has evolved over time to get intertwined in a broader perspective involving child welfare, governmental authority, parental rights and political agenda.

### 5.3.5. Content Analysis

The initial narrative in the beginning of the topic is largely accusatory, indicating a public perception that Socialtjänsten might be acting unjustly or based on misinformation. The sentiment expressed range from simple accusation to intense emotional outbursts. However, this period also included a subset of posts that questioned the accusatory narrative, suggesting that such claims may be born out of misinformation or misunderstanding. A few others defended Socialtjänsten's actions by arguing that child removal procedures were carried out for the child's best interest, further adding diversity to the narrative.

Following this period, a seven-month span from March 31 to September 16, 2022, saw a continuation of the discussion, but in a much less frequent manner. Some posts during this period indicate a shift in public discourse, referring to the events of February 2022. This period also saw a novel approach to the narrative with several posts employing sarcasm or satire to address the accusations of child kidnapping, possibly aiming to dismiss or discredit these claims.

In the final period another rapid increase in discussion could be observed, which continued for the next three upcoming months and then decreased. The first three months indicated a resurgence of the narrative. This period was marked by a strong division in public opinion and overall increase in negative sentiment. At the end of the final period some attempts were made to redefine the term “kidnapping” to add nuances to the discussion.

In conclusion, the narrative around allegations of Socialtjänsten and child kidnappings evolved over time, marked by periods of intense discussions and shifts in public sentiment. Despite periods of relative calm, the narrative seems to resurface, with people having strong opinions and polarising further apart from each other based on the sentiment in Figure 5.

## 5.4. Topic 77

Topic 77 referred to as childr - kidnappa - kidnapping - families - kidnapped - sweden - diab - kidnap - parents - talal by BERTopic consist of 97 posts. A detailed examination of the data reveals that the discussions mainly revolve around alleged instances of child kidnapping by Socialtjänsten. In total the number of posts that were classified as potential disinformation is making up 0.001% of the entire topic data. The average number of posts per month for the given topic is 8.25 with sudden spikes in Februari, November, December 2022 and January 2023, presented in Figure 7.

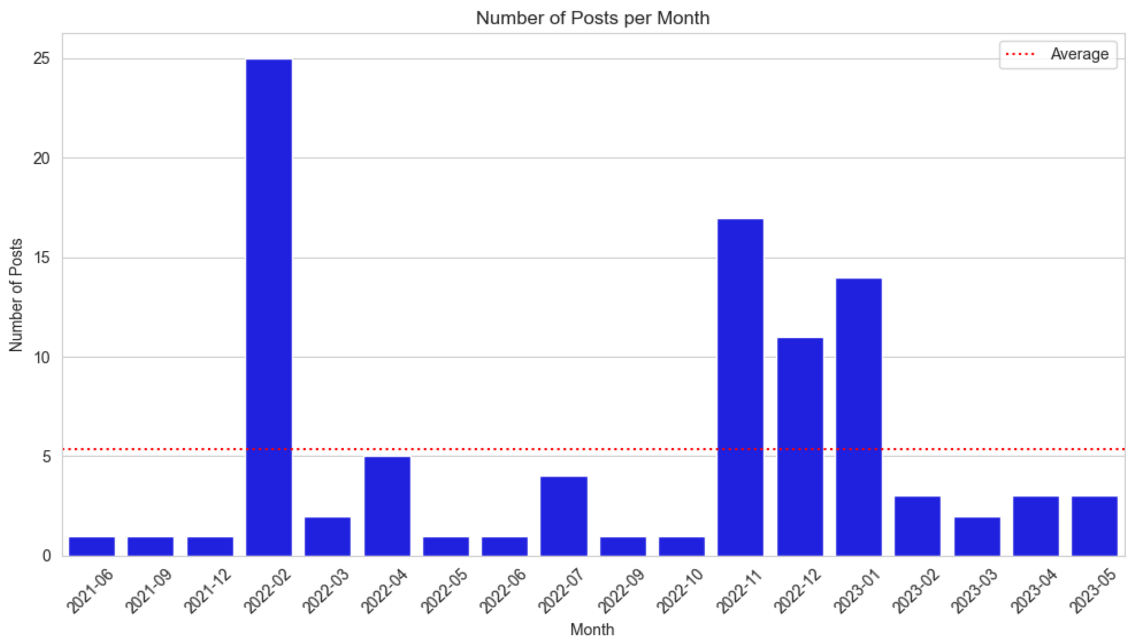


*Fig 7. Distribution of posts for topic 77*

### 5.4.1. Representative Topic Posts

To gain a better understanding of the topic, a couple of carefully selected data points are presented below. Accompanying each post is the date it was posted, the probability of it belonging to the topic, if the post is representative of the topic and its sentiment score. The rolling average sentiment over time for topic 77 can be seen in Figure 8.

| **Date** | **Probability** | **Representative** | **Sentiment score** |
|---|---|---|---|
| 2022-02-08 | 1.0 | False | 0.0 |

"Facility in Sweden called Socialstyrelsen kidnapping the children against their will infront of their parents without mercy and they force them to live with stranger families . All of this happens in civilised democrat country hypocritical State CIJ_ICJ UN hrw amnesty

| **Date** | **Probability** | **Representative** | **Sentiment score** |
|---|---|---|---|
| 2022-02-11 | 1.0 | False | -0.83 |

"Tycker det är väldigt obehagligt hur det nu trummas ut att svenska myndigheter kidnappar barn . Efter några år i förvaltningsrätten , som nämndeman , kan jag säga att ingen tar något barn från föräldrarna lättvindigt . Oroar mig för den
här utvecklingen och hoten mot socialtjänsten

| **Date** | **Probability** | **Representative** | **Sentiment score** |
|---|---|---|---|
| 2022-11-26 | 1.0 | False | -0.82 |

”Sverige har blivit ett fascist land . Socialtjänsten som kidnappar barn med polisen som hantlangare . Rättssystemet som tar deras linje per automatik . Detta strider mot demokrati och mänskliga rättigheter

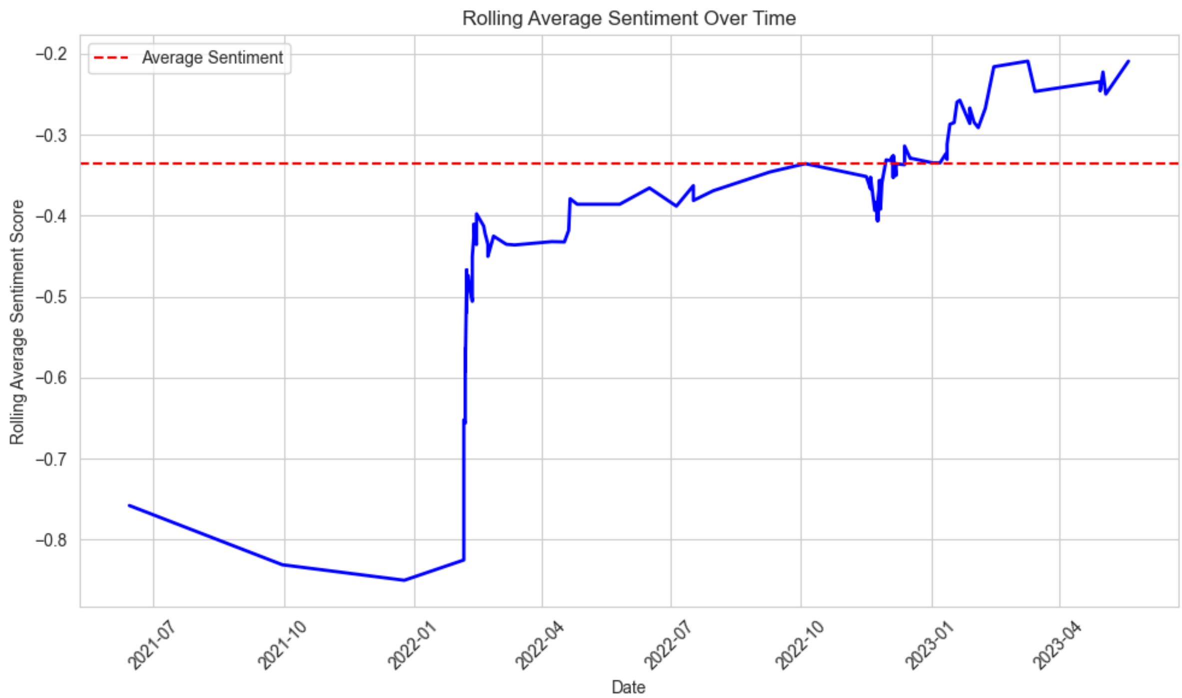


*Fig 8. Rolling average sentiment over time for topic 77*

### 5.4.2. Period 2022-02-05 To 2022-02-22

During the period in february 2022 several posts accuse Socialstjänsten of engaging in child kidnapping, targeting various ethnicities, including Syrians, Somalians as well as Swedish children. Throughout the discussion claims of unjustified forceful removal of children based on unfounded accusations and lies by Socialtjänsten. Criticism is raised against how Socialtjänsten dismisses the accusations without providing evidence or investigations, it is also emphasised that existing videos and testimonies from families serve as potential proof of Socialtjänstens wrongdoings and that the agency is trying to hide and deny the truth. It is also mentioned that Socialtjänsten is exploiting immigrants for personal gain, like selling them or giving them to families who do not have children. The following posts have been chosen to represent the discourse of the period.

| **Date** | **Probability** | **Representative** | **Sentiment score** |
|---|---|---|---|
| 2022-02-05 | 0.73 | False | 0.0 |

"The Sweden Socialtjänsten kidnapping the kids from their families by law and force with a ridiculous reasons . Hundred of Syrian , Somalian and other nationalities children are been kidnapped and separated from their families every day . This brutality should stop immediately . "

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-06 | 1.0 | False | -0.81 |

"Det må hända finns fall då socialtjänsten gjort rätt , men vi är många svenskar som i åratal skrikit oss hesa om att socialtjänsten "" kidnappar "" våra barn på rättsosäkra grunder . Grunderna är ofta förtal utan bevis och uppenbara lögner . Varför är det helt plötsligt desinformation ? "

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-07 | 0.58 | False | -0.36 |

" Detta är ingen hotkampanj era lögnare , utan det är sanningen som ni försöker förneka och dölja ! Socialtjänsten i Sverige har i många år kidnappat en hel del barn av olika svepskäl ! Stop_kidnapping_our_children “

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-02-14 | 0.55 | False | 0.0 |

"socialstyrelsen They open their door for them , it becomes clear now ! There is something hidden for kidnapping children one of them to give that children to Sweden families that don t have kids and give them money for , the second is to Exploiting immigrants coz they know that there is no power "

### 5.4.3. Period 2022-11-24 To 2023-01-31

The period through november, december and january continues discussion of accusations against Socialtjänstens involvement in child kidnapping. Various accusations and negative sentiments were expressed towards Socialtjänsten, where claims that the government was deliberately stealing children and engaging in child trafficking were expressed. However, a segment of the period mentions Gothenburg's actions against the rumours being spread about Socialtjänsten, where Socisaltjänsten in Gothenburg have reported the issue to the police. This handling of the situation received support from some individuals who expressed appreciation of Gothenburg’s efforts to address the situation. Nonetheless, the period was dominated by negative themes surrounding Socialtjänsten. The following posts have been chosen to represent the discourse of the period.

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2022-11-24 | 0.81 | False | 0.25 |

¶"I sociala medier sprids bilden att svensk socialtjänst kidnappar barn och nu delas också en video där attentat mot socialtjänsten nämns . Nu väljer därförsocialtjänsten i Göteborg att göra ett utskick till sina medarbetare och att göra en polisanmälan "

| Date | Probability | Representative | Sentiment score |
|---|---|---|---|
| 2023-01-02 | 0.99 | False | -0.31 |

¶" Detta är ingen myt eller desinformationskampanj . Socialtjänsten kidnappar barn . Politiker tjänar grova pengar på denna barnhandel Sweden Norway "

### 5.4.4. Content Analysis

There are several different narratives present in topic 77, the dominating narrative throughout the topic are accusations of child kidnapping with calls to immediately halt these unfounded actions. These kidnappings are focused mainly around children of muslim background, but children with Swedish background are also mentioned. Notable is the focus on how Socialtjänsten has handled the rumours of unlawful kidnappings, with claims that they have presented no proof of their stance or done any investigations on the issue. One narrative in topic 77 focuses on the way Gothenburg handled the rumours being spread about Gothenburg's Socialtjänst, in the data in topic 77 these actions are met with overwhelmingly supportive reception. However, the overall narrative in topic 77 maintains the negative theme surrounding Socialtjänsten. Even though the majority of discussion is negative, it is important to note that not all the posts are in favour of these narratives speaking against Socialtjänsten.

# 6. Discussion

The following section revolves around a detailed investigation into prevalent misinformation narratives related to Socialtjänsten, employing several analysis tools like VADER and BERTopic to assess sentiment and topical significance in online discourse. It also discusses the possibility of applying these methodologies beyond their current scope for continuous monitoring, misinformation detection, and extended research in other domains.

## 6.1. Narrative Discussion

When conducting the analysis of the interesting topics extracted from the data we identified seven primary potential misinformation narratives that were consistently prevalent across the three topics. These narratives were chosen because of their distinct perspectives and negative sentiments around Socialtjänsten.

The first prevalent narrative identified asserted that “Socialtjänsten is involved in corrupt processes”, implying that the institution is acting on unlawful grounds within the system. The second narrative suggested that “Socialtjänsten takes away children without good reason”, which was a major perspective throughout the discourse in the topics. This narrative lays ground to fear of unjustified state intervention in family matters. The fifth primary narrative argued that “Socialtjänstens’ workers are unqualified”, casting doubt on the professional competence of those working within the institution. The sixth narrative implies that “Socialtjänsten is unaccountable” meaning a lack of responsibility and transparency. This is referring to claims that Socialtjänstens does not provide any proof or perform any investigations to strengthen their position.

The third narrative claimed that “Socialtjänsten discriminates against certain demographic groups”, suggesting unfair treatment based on factors such as race and religion. There are many examples of this throughout the topics, where many claims were made that Socialtjänsten specifically targets and kidnaps Muslim children. The fourth narrative implies that “Socialtjänsten places Muslim children in Christian homes”, suggesting that one of the institution’s actions is to Christianise the children. The seventh and last primary narrative we identified suggests that “Socialtjänsten is a part of a larger governmental conspiracy”, implying that the institution is involved in sinister, large-scale state operations. Examples of this include some posts claiming that the kidnapped children are given to Swedish families without children or that the institution is selling the children.

## 6.2. VADER

Throughout our analysis, the use of VADER facilitated an in-depth exploration of the sentiments associated with the discourse surrounding the Socialstyrelsen. This sentiment analysis tool offered support in terms of assessing the intensity of discussions, tracking shifts in emotional intensity over time. Each time a significant event or announcement led to a surge in discourse about Socialstyrelsen, we could observe a corresponding spike in sentiment intensity, highlighting the reactive nature of public sentiment in response to unfolding developments.

However, despite its usefulness, we must exercise caution when interpreting the results produced by VADER. Like all analytical tools, it has its limitations. Its rule-based model may potentially prime us to falsely interpret the emotion of a certain post. As it operates based on a predefined sentiment lexicon and set rules, it may occasionally struggle to accurately capture the sentiment of texts with complex or nuanced expressions, cultural differences in language use, or context-specific language. Moreover, VADER primarily looks at the immediate linguistic context of a given word or phrase irrespective of prior post connected to the response. Consequently it may overlook broader contextual cues, which can sometimes lead to misinterpretation of sentiment.

While VADER provided valuable insights into our project, it doesn't paint the complete picture. The intricacies of sentiment cannot be fully encapsulated through a single tool, emphasising the need for a multi-pronged approach in sentiment analysis. VADER can certainly serve as a powerful tool when used in conjunction with other analytical methods. In particular, Bitext Mining may complement the utility of VADER, as it is an effective tool for relation extraction, or identifying relationships and associations between entities in the text. This can enhance the analysis by providing additional context to the sentiment scores generated by VADER. For instance, it can help discern whether the sentiment is directed towards Socialstyrelsen as a whole, specific services, or individual actors within the institution.

Conclusively, our study is an example of the utilisation of Vader in sentiment analysis, specifically in tracking change of intensity of sentiment over time, as to identify shifts in discourse. Our approach underscores the need for a more holistic approach and the desire to integrate multiple tools and techniques to gain a more comprehensive understanding on public opinions in online discourse, such an approach would enhance the robustness of sentiment analysis, equipping it to navigate the complex landscape of online discourse more efficiently

## 6.3. BERTopic

Bertopic produced 619 different topics. This large amount of topics helped us identify several thematically similar narratives we might otherwise have missed. The large number of topics is a result of the usage of c-TF-IDF topic representation as it has the ability to differentiate similar topics especially based on rarer n-grams such as “Diab Al Talal”, which caused the split between topic 52 and 77. The drawback of this feature is the relatively small size of some topics such as topic 492 which only contains 15 posts in total. However the topics we found to be interesting were all larger, so the opposing view to this is that it helped remove potentially uninteresting data from otherwise interesting topics. Unfortunately the amount of manual analysing required to go through several hundred topics is time consuming enough not to be possible. However given more time an automated approach of analysing the topics could be developed to solve this flaw.

Possible effect of removing duplicates: as several of the posts that are manually identifiable as misinformation are made up of exactly the same text, this means we’ve possibly removed quite a large amount of the exact data we were hoping to analyse.

## 6.4. Future Implications

The study and its methods present a robust approach to information extraction and analysis from online social media platforms, an area that is becoming increasingly crucial as digital communication continues to permeate our society. In a practical context, the methods used in this project—web scraping, data preprocessing, sentiment analysis with VADER, and topic modelling—can have broad applications. They can be utilised in various contexts, such as market research, political discourse analysis, and public opinion monitoring on social issues. In the future, the project could be extended to include more social media platforms, and different forms of digital communication like blogs, news websites, and forums. By broadening the scope, a more comprehensive picture of online discourse could be achieved. The effectiveness of the analysis could potentially be improved by incorporating more advanced techniques. In addition to the VADER and BERTopic models used, other machine learning models could be explored to better capture the subtleties of sentiment and themes in the data.

Web scraping is a pivotal part of this project, and future research could focus on developing more advanced and ethical scraping techniques that ensure data quality and respect users' privacy. The project also offers insights into misinformation detection. The spread of misinformation on social platforms could be effectively tracked and analysed using the methods applied in this project. This could potentially lead to the development of more sophisticated tools for misinformation detection and prevention in the future.

Lastly, ethical implications of this research should be given due consideration. Ensuring anonymization of data and respecting the privacy of individuals whose posts are being analysed are critical. Future research could also probe the ethical implications of using AI for sentiment analysis and topic modelling, such as potential bias in the algorithms. Reflecting on the potential impact and applications of our project results, we can envisage several areas where our work can be beneficial to KAPI, Socialtjänsten, Socialstyrelsen, and other organisations.

### 6.4.1. Direct Applications For Decision Making

Our project could contribute to informed decision making concerning communication with the intent to hinder and prevent the spread of rumours and disinformation. By scraping large amounts of data and using our classifier and topic modelling analysis, we can identify online disinformation related to Socialtjänsten. Once we gathered and analysed the data, we provided it to KAPI, which is responsible for submitting a report to Socialstyrelsen. Our results could enable KAPI to write a report that accurately portrays current online trends of misinformation, describing their content, and the channels in which the information is being spread. This data-driven approach could allow KAPI to provide improved advice to Socialstyrelsen on their communication efforts, both immediate and long-term.

### 6.4.2. Continuous Monitoring And Analysis

A long-term objective of our project is to develop a semi-automated service package that we can release to KAPI for their own use. This service package would include scraping scripts, data cleaner(s) for refining scraped data, analytic tools for understanding narratives within the data, and a packaging of these components into a cohesive tool. This service could potentially transform the way organisations like KAPI monitor online discussions. By streamlining the process of gathering and analysing online data, these tools could save significant time and resources. Moreover, continuous monitoring could help these institutions stay informed and proactive in addressing emerging concerns and misinformation.

### 6.4.3. Extended Research And Analysis

The methods and tools we developed for this project could potentially be applied to other contexts beyond Socialtjänseten. The scraping scripts, for instance, can be adapted for other online sites, applications, or topics of interest. With some modifications, these tools could be used in domains such as market research, political science, and public health. They could be adapted to monitor and analyse public sentiment regarding a specific policy or candidate in the political realm, potentially enabling a more comprehensive understanding of public opinion.

In conclusion, our project could potentially offer several avenues for direct application, continuous monitoring, and extended research. As we continue to refine these tools, they could play a role in ongoing efforts to counteract misinformation and enhance public understanding of important institutions and issues.

The amount of data could have been considerably larger and more diverse. Meaning more websites could have been scraped, such as Instagram, Youtube and different news sites where discussion of the topic investigated in this project are present. A larger dataset can provide a more accurate representation of the phenomenon being studied. With more data, the analysis is generally less likely to be influenced by random outliers, which results in more reliable and robust topics. With a larger sample size the ability to detect significant patterns or trends might be improved. Including data from various social media sites could provide valuable insights into public opinions and expert viewpoints. Analysing discussions and sentiments from these sources would have enriched the understanding of how the topic is being portrayed and discussed in mainstream media. A more extensive dataset could also have facilitated correlations within specific demographic segments or user categories. Finally a larger dataset would have increased the statistical power of the analysis, meaning a more comprehensive and accurate assessment of the research topic, increasing the confidence in the findings. It is however important to note that while having more data is generally advantageous, an excessively large dataset can have potential drawbacks. Other than obvious resource intensive drawbacks a large dataset could increase potential noise and irrelevance within the dataset. If the dataset includes unrelated or low quality information it can increase the risk of inaccuracies and irrelevant topics.

### 6.4.4. Future Of Fake News Research

There is no reason to believe that fake news will become less of a problem than it already is, especially considering today's volatile state of the world. And as previously mentioned, literature on modern day fake news on social media is scarce and very short-term based. There are no significant long term based effects reported as of now. Former Twitter CEO Jack Dorsey said in an interview that the problem with fake news on Twitter is that there is no one fix that completely solves the problem since there are too many elements that fake news consists of (The Indian Express, 2018). However, he considered the use of AI to bring the solution that he calls near perfect. The problem with letting human-developed AI be the gatekeeper of what can be considered fake news is that it is in turn based on human biases. Even though we never reached the stage where we would create our own fake news classifier, the discussions surrounding the topic often circulate around who is to determine what is fake news or not. Future researchers surrounding fake news are guaranteed to have to find solutions or ways around this problem.

Researching for this report, several papers gave suggestions to how one could create AI that solves this issue, or at least gives humans the option to give opinions on whether the classification of fake news is correct. Two students from Linköping University did their bachelor's thesis on fake news and COVID-19 found that the language used when discussing conspiracies relating to the pandemic, was more volatile than that of non-conspiratorial posts on Reddit (Savinainen & Hansen, 2022). If this is a proven significant effect on all platforms, it suggests that you can use natural language processing algorithms to flag suspected fake news when the language used can be classified as volatile. The classification of volatile language can be seen as unbiased when using sentiment analysis on texts, and the flagging system leaves the classification of fake news up to a human or other types of AI that can be used in tandem with the natural language processing algorithms.

Another paper attempts to create a clear cut fake news classifier, where it classifies texts as either proved with sufficient evidence or classified as not enough evidence (Atanasova et. al, 2022). This is done by giving a natural language processing AI a training dataset to learn how sufficient or non-sufficient claims are structured. A claim could be "By April 9, less than 9000 who tested positive for covid-19 in the UK died of the virus". The classifier would then mark the claim as supported if the facts say that 8958 people died of covid by April 9, as the evidence matched the claim. If you can train a classifier with enough data where you have a claim and a source with a classification of supported or not supported, you could match it with a database of evidence to successfully create an algorithm that with certainty can search and remove fake news posts. This gives hope to a future where social media can be successfully monitored by nun-human programs, since the amount of data that has to be checked simply isn’t possible without the algorithms. With the increasing strength of AI, one could argue that fact checking algorithms are the future, and that the future is bright. However there is an argument to be made that increased AI strength does nothing, as the same AI can be used to improve fake news instead. Since fact checking models and its use by social media companies are virtually unstudied, there is plenty of work to be done to discover whether it is a viable path to solving the problem.

# 7. Conclusions

In conclusion, our data-driven approach involving web scraping, sentiment analysis through VADER and topic modelling forged by BERTopic has provided novel insights into the dissemination of misinformation surrounding Socialtjänsten. This methodical exploration of online discourse has not only shed light on the channels through which misinformation is being spread but also presented an existence of a negative picture of public sentiment towards Socialstyrelsen. Through the identification of narratives inside topics together with sentiment analysis, and the use of Swedish Defence Universities paper we were able to identify the widespread extent of the misinformation and the attempt to undermine or erode the institutional credibility.

While the study's implications concern detection of misinformation about Socialstyrelsen, the methods employed offer broad applicability in other areas as well. The same data-driven approach can be employed for different institutions, subject matters, and even geographic locations, providing valuable insights to policymakers, researchers, and public administration officials. Furthermore, they can inform the development of more sophisticated tools for misinformation detection and sentiment analysis, thereby contributing to the advancement of methods and techniques within the field of misinformation studies.

# References

## Articles


Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2022). Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, *10*, 746–763. https://doi.org/10.1162/tacl_a_00486

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 160–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Diouf, R., Sarr, E. N., Sall, O., Birregah, B., Bousso, M., & Mbaye, S. N. (2019). Web scraping: State-of-the-art and areas of application. *2019 IEEE International Conference on Big Data (Big Data)*, 6040–6042. https://doi.org/10.1109/BigData47090.2019.9005594

Fedorenko, V., Fedorenko, M. (2022). Russia's military invasion of Ukraine in 2022: Aim, reasons, and implications. *Krytyka Prawa. Niezależne Studia Nad Prawem, 14*(1), 7–42. https://doi.org/10.7206/kp.2080-1084.506

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. (Working Paper) *arXiv*. https://doi.org/10.48550/arXiv.2203.05794

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modelling: Models, applications, a survey. *Multimedia Tools and Applications, 78,* 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Krotov, V., & Silva, L. (2018). Legality and ethics of web scraping. *Twenty-fourth Americas Conference on Information Systems*, 1–5. https://www.researchgate.net/publication/324907302_Legality_and_Ethics_of_Web_Scraping

Linvill, D. L., Boatwright, B. C., Grant, W. J., & Warren, P. L. (2019). "The Russians are hacking my brain!" Investigating Russia's internet research agency twitter tactics during the 2016 United States presidential campaign. *Computers in Human Behaviour, 99,* 292–300. https://doi.org/10.1016/j.chb.2019.05.027

McKay, S., Tenove, C. (2021). Disinformation as a threat to deliberative democracy. *Political Research Quarterly, 74*(3), 703–717. https://doi.org/10.1177/1065912920938143

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv.* https://doi.org/10.48550/arXiv.1802.03426

Pramana, R., Debora, Subroto, J. J., Gunawan, A. A. S., & Anderies. (2022). Systematic literature review of stemming and lemmatization performance for sentence similarity. *IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA),* 1–6. https://doi.org/10.1109/ICITDA55840.2022.9971451

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *Association for Computational Linguistics,* 3982–3992. https://doi.org/10.18653/v1/D19-1410

Stachofsky, J., Schaupp, L. C., & Crossler, R. E., (2023). Measuring the effect of political alignment, platforms, and fake news consumption on voter concern for election processes. *Government Information Quarterly, 101810.* https://doi.org/10.1016/j.giq.2023.101810

Udupa, A., Adarsh, K. N., Aravinda, A., Godihal, N. H., & Kayrvizhy, N. (2022). An exploratory analysis of GDSMM and BERTopic on short text topic modelling. *2022 Fourth International Conference on Cognitive Computing and Information Processing (CCIP),* 1–9. https://doi.org/10.1109/CCIP57447.2022.10058687

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems, 94,* 101582. https://doi.org/10.1016/j.is.2020.101582

Zou, F. (2022). Research on data cleaning in a big data environment. *2022 International Conference on Cloud Computing, Big Data and Internet of Things (3CBIT),* 145–148. https://doi.org/10.1109/3CBIT57391.2022.00037

## Reports and Bachelor’s theses

Ranstorp, M., & Ahlerup, L. (2023). *LVU-kampanjen: Desinformation, konspirationsteorier, och kopplingarna mellan det inhemska och det internationella i relation till informationspåverkan från icke-statliga aktörer*. Försvarshögskolan. ISBN: 978-91-88975-28-7 https://www.fhs.se/download/18.32d29dd2187bd01d5e455265/1682576119173/LVU-kampanjen.pdf

Savinainen, O., & Hansen, T. H. (2022). *Covid-19 related conspiracy theories on social media: How to identify misinformation through patterns in language usage on social media*. [Bachelor’s thesis, Linköping University]. Diva portal. https://www.diva-portal.org/smash/get/diva2:1679954/FULLTEXT01.pdf

## Websites

BeautifulSoup. (2023). *Beautiful soup documentation*. Crummy. https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Federal Bureau of Investigation. (July 13, 2018). *Russian interference in 2016 U.S. elections.* https://www.fbi.gov/wanted/cyber/russian-interference-in-2016-u-s-elections

Google trends (n.d) “*Fake news”,* Retrieved April 29, 2023. https://trends.google.com/trends/explore?date=all&q=%22fake%20news%22

Hamadé, K. (March 9, 2023). *Husein Miftar, 62, hittad död i häktet*. Expressen. https://www.expressen.se/nyheter/husein-miftar-62-hittad-dod-i-haktet-/

Nationalencyklopedin. (March 24, 2023). *TikTok.* https://www.ne.se/uppslagsverk/encyklopedi/l%C3%A5ng/tiktok

Nordevik, A. (November 28, 2022). *Extremistkampanj mot socialtjänsten tillbaka -"Kommer inte försvinna"*. Dagens Samhälle . https://www.dagenssamhalle.se/samhalle-och-valfard/socialtjanst/extremistkampanj-mot-socialtjansten-tillbaka-kommer-inte-forsvinna/

Partiet Nyans. (January 25, 2020). *Om partiet - partiet nyans*. https://www.partietnyans.se/om-partiet/

PTI Agency. (November 12, 2018). *Twitter CEO Jack Dorsey admits fake news problem but says no 'one fix' solution*. The Indian Express. https://indianexpress.com/article/technology/tech-news-technology/fake-news-a-multi-variable-problem-there-is-no-one-fix-twitter-ceo-jack-dorsey-5442759/

Schwartz, E. (February 7, 2022). *Socialarbetare: "Stor spridning på de här ryktena"*. SVT. https://www.svt.se/nyheter/inrikes/det-som-hant-har-varit-omtumlande

Shear, M. D. (March 15, 2017). *Trump calls 2005 tax return release 'fake news'*. The New York Times. https://www.nytimes.com/2017/03/15/us/politics/trump-calls-2005-tax-return-release-fake-news.html

Scrapy. (2023). *Scrapy 2.5 documentation*. https://docs.scrapy.org/en/latest/

Selenium. (2023). *Selenium with Python*. https://selenium-python.readthedocs.io/