TURUN
YLIOPISTO

Crisis Narratives and Topics of the COVID-19 Pandemic on Finnish Twitter

Otto Tarkka
Master's Thesis
Language Specialist Degree Programme, Digital Language Studies
School of Languages and Translation Studies
Faculty of Humanities
University of Turku
August 2022

Master's Thesis
Language Specialist Degree Programme, Digital Language Studies
Otto Tarkka
Crisis Narratives and Topics of the COVID-19 Pandemic on Finnish Twitter
60 pages, 2 appendices

Narratives are central to the human experience because they inform how we make sense of the complex world around us. When the COVID-19 pandemic forced people to stay home and avoid physical contact, narratives of the evolving crisis were actively told on Twitter. In this thesis, I analyse crisis narratives and their development on Finnish Twitter. I use a dataset of 375,322 tweets that were collected between January 2020 and August 2021. The data are analysed with the help of topic modelling, a machine-learning method that is used to discover latent topics from large collections of texts. The most common topics and their temporal distribution as well as tweets that are strongly associated with these topics are analysed within the theoretical framework of crisis and narrative studies. The results show that the most common topics were related to the measures put in place to control the spread of the virus. The COVID-19 pandemic appeared to many as a crisis of regulations rather than as a health crisis. As relatively few people were affected by the virus itself, the crisis narratives shared on Twitter were more concerned with the impact the everchanging restrictions and guidelines had on people's everyday lives than the foreign threat imposed by the virus. These results provide insight into the ways in which crises are constructed in narrative and thus can be used to better understand how future crises emerge and evolve.

Keywords: Twitter, crisis narratives, topic modelling, LDA, COVID-19

Contents

1 Introduction

When looking back at the COVID-19 pandemic that began as a small local outbreak in China in late 2019 and developed into a global pandemic in 2020, a myriad of memories is awakened in one's mind. Personally, I recall moments of fear, grief, and uncertainty, but also of joy, hope and, excitement. I believe that many can relate to these feelings, and that we all have our own story to tell about how we experienced the pandemic. And not just one story, but many stories; some smaller, some greater. This thesis is about those narratives. Narratives are how we remember and make sense of the world in which we live (Heath, 2013). To understand how the COVID-19 pandemic took place, beyond simple infection and death statistics, one must understand the narratives that were told, are being told and will be told about it. Certainly, there will be many papers and books written about the COVID-19 crisis in the coming years and decades. This thesis aims to be part of that discourse.

The COVID-19 pandemic started in Wuhan, China in late 2019 and spread to mainland Europe and beyond in early 2020. On 11 March 2020, the World Health Organisation declared the worsening situation a pandemic (WHO, 2020b). In response, nations around the globe began closing their borders and limiting the movement of their citizens in order to halt the spread of the virus. In the age of social media, people turned to Twitter and other social networking sites to offer their opinions and to look for answers and comfort. The WHO published a statement calling the situation an "'infodemic' – an over-abundance of information – some accurate and some not" and announced that they would be working together with social media companies to respond to myths and rumours (WHO, 2020a). WHO's statement made it clear, that the COVID-19 pandemic would be much more than just a health issue.

A human is a story-telling animal, who likes to share their experiences with others. As Ochs and Capps (2001, p. 1) put it: "[W]hen people are together, they are inclined to talk about events – those they have heard or read about, those they have experienced directly, and those they imagine." Since coming together physically was discouraged or prohibited during the COVID-19 crisis, people came together on online social platforms. The first global pandemic in the age of social media lead to a situation in which information was created, shared and updated at a pace no individual could follow. A multitude of crisis narratives emerged both in traditional media and social media. The quickly evolving pandemic and rapid flow of new information led to fake news, questioning of health care experts, and the development of alternative expertise on Twitter (Au & Eyal, 2022; Väliverronen et al., 2020; van Dijck & Alinead, 2020).

Understanding peoples' reactions to global crises, like the COVID-19 pandemic, is crucial, if we wish to mitigate the worst effects of fake news, disinformation, and conspiracy theories. We must understand how crisis narratives are created and challenged, and what actors wield the greatest power in the creation of these narratives. We live in an age of crisis, where financial crises are followed by health crises, followed by war. All the while, climate change poses an ever-present existential threat to mankind. Climate change is predicted to increase the occurrence of crises such as extreme weather events (Stott, 2016) and number of health related risks (Haines & Ebi, 2019), which is why the COVID-19 pandemic should be taken as a learning experience to avoid the mistakes made by governments and authorities in the handling of the crisis. Narratives and the world they describe are cocreated synchronously, both reacting to changes in the other. Therefore, language can be employed as a tool to control the manner in which crisis evolves. On the other hand, language also has the power to exacerbate the negative outcomes of crisis. In order to manage and predict how future crises might unfold, a thorough understanding of the language of crisis and the formation of crisis narratives is paramount.

In this thesis, I examine the COVID-19 pandemic as a crisis that not only exists in the material realm but is also created through discourse and narratives. Discourse in this thesis is seen as linguistic social interaction between participants. As Hay (1996, p. 255) argues: "[A crisis] is subjectively perceived and hence brought into existence through narrative and discourse". In this view, without discourse, there would be no crisis. COVID-19 is not only a pandemic, but also a linguistic event with its own linguistic characteristics. Therefore, in this thesis, I analyse Finnish messages sent out on the social media platform Twitter (known as *tweets)* to see how crisis narratives are constructed, affected by, and reacted to in public digital discourse. My research questions are as follows:

(1) Who were the most active participants in creating crisis narratives on Twitter?

(2) What were the most common hashtags and how do they relate to the crisis narrative?

(3) What topics and narratives within those topics arose in Finnish Twitter discussions during the COVID-19 pandemic?

(4) How did the distribution of topics and the crisis narratives change during the course of the pandemic?

To answer these research questions, I analyse Finnish tweets related to the COVID-19 pandemic using topic modelling. The functioning and implementation of the topic model will be discussed in detail in sections 2.5 and 3.3. Based on the results of the topic modelling, I perform a quantitative and qualitative analysis of the most prevalent COVID-19 related topics to see how crisis discourse and crisis narratives emerged and developed over the course of more than 1.5 years. Additionally, a brief analysis of the most active tweeters and most common hashtags is conducted.

The main motivation behind the chosen method is the scale of the data. The dataset used in this study consists of nearly 400,000 tweets, which in turn contain more than 8 million words. Computational methods are a must for adequately analysing a dataset of this size, since it cannot possibly be fully comprehended and analysed by a single human or even group of people. Topic modelling is a computational method that is used to discover latent topics from a large collection of texts. It represents a these texts as a set of topics, which can be understood and further analysed by a human researcher (Jacobs & Tschötschel, 2019). However, the computational method used in this study is only a means to an end. The topic model cannot answer any question on its own. Therefore, a thorough qualitative analysis is required if adequate insight into the subject matter is desired. This central tenet of corpus-assisted discourse studies (CADS) is put well in Partington et al. (2013):

> By combining the quantitative approach, that is, statistical overviews of large amounts of the discourse in question – more precisely, large numbers of tokens of the discourse type under study contained in a corpus – with the more qualitative approach typical of discourse analysis, that is, the close, detailed analysis of particular stretches of discourse – stretches whose particularly interesting nature may well have been identified by the initial overview – it may be possible to better understand the processes at play in the discourse type (p. 11).

This thesis is structured as follows: First, I give an overview of the development of the COVID-19 pandemic in Finland. Then, I introduce theoretical perspectives on crises and the telling of narratives. The fields of Digital Discourse Analysis, Corpus-Assisted Discourse Studies and the functioning of topic models are also discussed. Then, I discuss Twitter as a platform and the typical characteristics of language on Twitter. In the material and methods section, I present the data and methodology used in this research. Then, I analyse and discuss the results of the study and finally draw conclusions from the results and what they mean for future research.

This study is multidisciplinary by nature. I draw from multiple fields of linguistics, namely discourse analysis, corpus studies, sociolinguistics, narrative studies, and natural language processing. I believe this broad approach is best suited for analysing a subject that penetrates and has radically altered many aspects of society and for working with a dataset of this magnitude.

2 Background and Theory

In this section, I first give a brief overview of the development of the COVID-19 pandemic in Finland from early 2020 to the summer of 2021. Then, I discuss how crises can be conceptualised and how they have been studied in the past. I then discuss narrative studies, especially the study of small stories and how they relate to the current study. I introduce the fields of linguistics known as Digital Discourse Analysis and Corpus-Assisted Discourse Studies. Finally, I explore some of the characteristics of language in Twitter messages.

2.1 Development of the COVID-19 pandemic in Finland

Before a comprehensive analysis of language of crisis narratives can be undertaken, a thorough understanding of the context in which the language was produced is crucial. Here, I provide an overview of the development of the COVID-19 situation in Finland and authorities' reactions to it. As with most other countries' governments, the Finnish government reacted to the pandemic by implementing restrictions on social gatherings. In many cases, the Finnish authorities decided to rely on recommendations rather than strict laws and restrictions to guide the population, relying on the good will and belief in a common good among the people. This approach was chosen partly because the authorities trusted the people to make the right choice and observe the recommendations and partly because signing restrictions into law proved to be difficult. For example, in March 2021 the government tried to impose strict restrictions on movement, but these measures were deemed unconstitutional. Restrictions and recommendations were given both on a national and a regional level and they were in constant fluctuation as the pandemic progressed, periodically worsening and improving. Hence, only a summary overview of the most significant changes in restrictions is provided here. Figure 1 shows the number and development of new infections by week in Finland from January 2020 to August 2021.

The first case of COVID-19 in Finland was reported on 29 January 2020 (Finnish Institute for Health and Welfare, 2020). This single infection did not lead to further infections and the situation remained calm, with only individual cases being reported in the month of February. In March, weekly new cases quickly rose to the hundreds, which prompted the government to enact the Emergency Powers act and impose strict restrictions on public gatherings and international travel (Finlex, 2020). Schools were also closed to halt the spread of the virus. On March 28, access to and from the region of Uusimaa, which was worst affected by the virus, to the rest of the country, was restricted (Finnish Government, 2020). In the first week of April, weekly reported cases peaked at 966 and restaurants were closed. As new cases slowly started to dwindle, access to and from the region of Uusimaa was reallowed. Schools reopened on May 14. In the summer of 2020, there were relatively few new reported cases and, consequently, restrictions were mostly lifted.

However, in late August and in September, more and more new cases were reported, and new restrictions were put in place. The second wave of the pandemic in Finland reached its apex in late November, when 3,136 new weekly cases were reported (Finnish Institute for Health and Welfare, 2021). The situation remained tense through the winter, although the newly approved vaccinations provided some hope for the future. The third wave of the pandemic peaked at 4,942 weekly cases at the beginning of March 2021 (Finnish Institute for Health and Welfare, 2021). This prompted the government to declare a state of emergency (Finnish Government, 2021a); municipal elections which were planned for April were postponed to June, restaurants were closed and even restrictions on movement were discussed, although eventually not put into law. In April and May 2021, new cases started to dwindle again and, with vaccination rates rising steadily, talk about an exit plan from a state of crisis back to normal began arising in the media and government (Finnish Government, 2021b).

Although, at the time of writing in the summer of 2022, the pandemic is not over yet and future developments remain unclear, the most acute stage crisis seems to be mostly behind us for now. With the background knowledge of how the virus spread and how it was fought in Finland, an analysis of Twitter messages from this period should give valuable insight into the crisis narratives that were relayed on the social media platform.

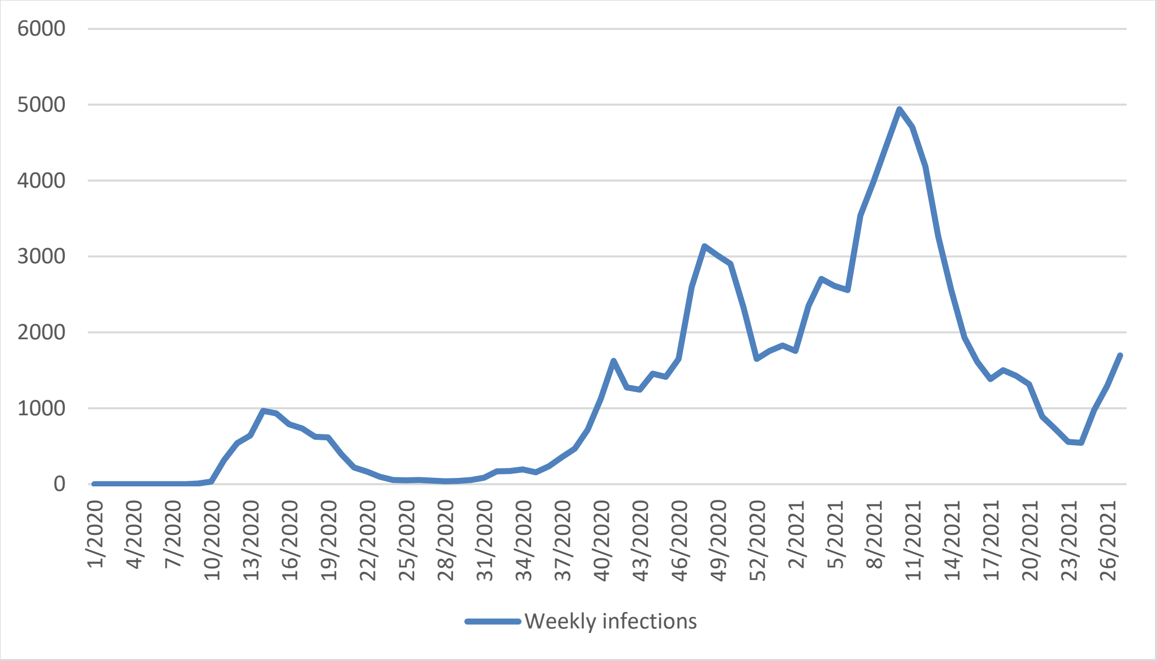


Figure 1. Number of weekly new confirmed COVID-19 infections (Finnish Institute for Health and Welfare, 2021)

2.2 Crises, narratives, and crisis narratives

Both crises and narratives are subjects that have been studied extensively. Yet, both areas of research are constantly developing as new theoretical frameworks emerge. In this section, I provide a review of some of the theoretical frameworks that have been developed regarding crises and narratives. Thereafter, I bring these two subjects together in a discussion about crisis narratives.

2.2.1 Crisis

Much of the research on crisis and crisis communication has been conducted from the perspective of crisis management and organizational response and preparedness (M. Huang, 2020). The field of study focusing on organizational response to crises is known as Social-Mediated Crisis Communication (SMCC) (Cheng & Cameron, 2017). In this thesis, however, I wish to approach crisis from a social perspective and explore how crisis is created and communicated by social agents in discourse. When viewed from this perspective, a crisis is never a purely objective thing that exists outside of the human experience but rather a socially constructed event. The COVID-19 pandemic can certainly be classified as a crisis, which has had profound impact on the material world. Moreover, and more importantly regarding this study, it is an event, that has been narratively constructed as a crisis. This crisis narrative has been present since the early stages of the pandemic, exemplified by the Finnish national broadcaster Yle, who first referred to the pandemic as a crisis ("koronakriisi") as early as March 2020.

Although there seems to be a general understanding of what the word 'crisis' means, a strict definition of the word is hard to pin down. Indeed, there are numerous ways to conceptualise what the word crisis means and much has been written about its nature. How one wishes to define the word depends partially on the approach and research question. De Rycker and Mohd Don (2013) provide an excellent overview of the study of crises and how the term has been defined by different scholars. Based on previous literature, De Rycker and Mohd Don list the following characteristics of a crisis often employed in the literature, among others: negative, disrupts an existing order, abnormal, sudden but not unexpected, involves extensive damage (pp. 8–9). All these characteristics can easily be applied to the COVID-19 crisis, too: The COVID-19 crisis has been overwhelmingly negative as it has led to millions of deaths and hospitalisations (WHO, 2021) as well as caused discomfort and anxiety in the form of regulations, such as lockdowns (Usher et al., 2020). These regulations have disrupted the daily lives of most people, forcing them to swiftly readjust their lifestyles. Also, the crisis started suddenly but not unexpectedly, as the virus had already spread in China and scientists knew a global pandemic was possible.

Attempts of formalizing the structure and development of crises have been made in crisis literature. This is an ambitious undertaking because a crisis is not an isolated, stagnant, or uniform event. It is ever evolving, "a moment and process of transformation" (Hay, 1996, p. 255). Fink (1986) proposes a four stage model for understanding the development of a crisis. In this model, first comes the prodromal crisis stage, which Fink describes as the warning stage before the crisis truly begins. If sufficient action is not taken during this stage, it is followed by the acute crisis stage. This is "the point of no return" (p. 22) when the kettle boils over, damage is done, money is lost. Next comes the chronic crisis stage, which is when the slow healing process begins. Finally comes the crisis resolution stage, when the crisis is over and "the patient is well and whole again" (p. 23). The cyclical nature of crisis manifests as the resolution leads to new contradictions that eventually lead to a new crisis. This model is useful in conceptualising the development of a crisis. However, it suffers from being underdeveloped and overly general. As Fink himself puts it: "crises historically evolve in cyclical fashion … [which] makes it difficult to see where and when one crisis ends and another begins" (pp. 27-28). Each crisis is a unique event in time-space and in constant fluctuation, which makes formulating a cohesive and generalizable structure a difficult, if not impossible, task.

2.2.2 Narrative

The study of narratives is the study of the stories that humans tell each other and themselves about the nature of reality. There are many diverging and conflicting scholarly traditions of narratives. Early studies of narratives took a 'Labovian' approach, in which a narrative was seen as a monological and concise story with a beginning, middle and end, often retelling past events in chronological order (Labov, 1972). In his work, which consisted mainly of interviews of African Americans in the inner city, Labov formulated a fully formed narrative, which consisted of an abstract, orientation, complicating action, evaluation, resolution, and coda. This method of analysis greatly informed researchers in social sciences, who began exploring narrative as a method of analysis in the 1980s in what became to be known as the narrative turn (De Fina & Georgakopoulou, 2011). However, this strict view on narrative has led to the dismissal of 'small stories', non-archetypal narratives and descriptions of ongoing events or hypothetical futures (Georgakopoulou, 2006). Today, narrative research has mainly moved away from trying to classify texts as either belonging or not belonging to the genre of narratives and rather approaches narratives as a flexible and diverse category (Page, 2015).

Conversation analysts (CA) critique the Labovian approach for analysing narrative as an abstract unit removed from the context in which it was told. Instead, they argue, narrative like all other verbal communication, should be seen as talk-in-interaction. Narrative, when approached from a CA perspective is seen as emerging in interaction between participants, co-constructed by both the teller and their audience, who through their mannerisms, reactions and interjections could modify tellings of narratives (Goodwin, 1984). The view of narratives changes from being structured monologues to polytonal, ever evolving interaction events. Further analyses of narrative go beyond the spoken or written word, and instead see narrative as "a mode of thought" (De Fina & Georgakopoulou, 2011, p. 15). This approach sees narrative as a fundamental and integral component of the human cognitive makeup. Narratives are thus a way for humans to make sense of the world. Narratives inform us about our surroundings and change how to act in response. They invite us to notice patterns and shape our perspective on the world, or as Heath (2013) puts it:

> Narratives are a way of thinking, a way of ordering the events of the world which would otherwise seem unpredictable or incoherent. As people order the events of

> their world into meaningful patterns, those patterns— scripts and themes— express their values and guide their actions. However, not everyone thinks and acts in the same way. Viewed this way, narrative analysis assumes that people choose among competing stories that account for a given event. Thus, the question is often not whether narrative, but which narrative is best? (p. 171)

Ochs and Capps (2001) analyse narrative of personal experience, which small stories often are, as a set of five dimensions, these being tellership, tellability, embeddedness, linearity, and moral stance. Tellership refers to the person(s) present at the telling of a story and their role in forming the narrative. A prototypical example would be one person telling a story to a passive audience. It is also possible for a story to have many active tellers, who interact with and add to the story. Tellability refers to the extent to which a story consists of reportable, interesting, events that can be delivered rhetorically effectively. Some textual signs of low tellability are cut off sentences, revisions of initial claims and repetition. Embeddedness refers to the ways and extent to which the story is connected to surrounding discourse and broader social context. In the dimension of linearity, highly linear stories follow singular, temporally ordered and causally relevant paths, whereas stories of low linearity follow diverse, uncertain paths. Finally, moral stance shows what the teller(s) believe is “good or valuable and how one ought to live in the world” (Ochs & Capps, 2001, p. 45).

The study of small stories is a relatively recent shift in narrative analysis. Promoted especially by Georgakopoulou (see e.g., Georgakopoulou, 2006, 2017), small stories research aims to rid itself from the last remnants of Labovian tradition and focus on small, everyday storytelling that people engage in in social interaction. These stories are often short, unfinished, unstructured, and created through interaction rather than told by a single teller. Small stories represent “a gamut of under-represented narrative activities, such as tellings of ongoing events, future or hypothetical events, shared (known) events, but also allusions to tellings, deferrals of tellings, and refusals to tell” (Georgakopoulou, 2006, p. 123). A small story approach can shed light on the intricacies, individual variation, and counterexamples within a broader meta-narrative (Phoenix et al., 2010). Whereas grand narratives tend to depict a deceptively simplified and smooth view of the world, small stories expose the internal inconsistencies within them.

2.2.3 Crisis narrative

The focus of this study is on narratives of crisis and crisis as a narrative. Hay (1996) illustrates the power of crisis narratives in his paper on the ‘Winter of Discontent’ and shows how the events that took place were framed as a crisis by the British media. In the winter of 1978–1979, a number of British trade unions went on strike. These strikes were harshly criticized by the media and, Hay argues, a crisis narrative was formulated, which led to a massive drop in the popularity of the then governing Labour Party in favour of Margaret Thatcher's Conservative Party. Hay emphasises the role of the media in the narration of the crisis. He claims that a multitude of separate, semi-synchronous events were connected in media discourse as one larger crisis event. Figure 2 shows how Hay visualizes a crisis narrative is formed in the media by highlighting selected events and statistics which are employed to bolster the narrative of a crisis ravaging society.

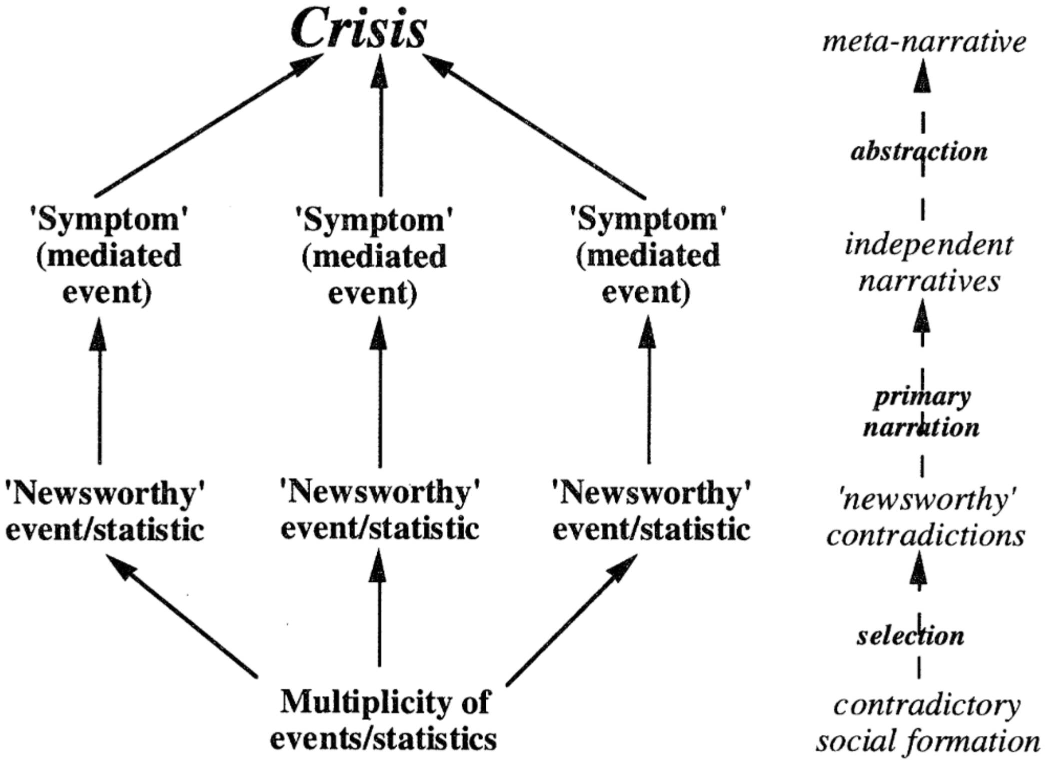

Crisis
meta-narrative
abstraction
'Symptom' (mediated event)
'Symptom' (mediated event)
'Symptom' (mediated event)
independent narratives
primary narration
'Newsworthy' event/statistic
'Newsworthy' event/statistic
'Newsworthy' event/statistic
'newsworthy' contradictions
selection
Multiplicity of events/statistics
contradictory social formation

Figure 2. Crisis as a meta-narrative from Hay, 1996. The figure shows how the media pick newsworthy events and statistics from a myriad of events and frame them as symptoms of the larger meta-narrative of a crisis.

To exemplify this in the context of the current study, we can imagine a great number of events and statistics relating to COVID-19: lethality, survival rate, daily new infections, super spreader events, lockdowns, restaurant closures, etc. Some of these events are deemed newsworthy, written about in traditional media and discussed on Twitter. As with the events of the Winter of Discontent, the most newsworthy events, and statistics (e.g., record high infection rates or new regulations) get more attention than others. The media and Twitter together amplify and spread these stories to reach a large audience. Alone, a single news headline or a tweet has little effect, but a nexus of hundreds or thousands of tweets together create the narrative of a crisis.

To further exemplify the interwoven nature of the material world and discourse in the building of a crisis is the ongoing discussion around climate change. In recent years, climate change and its effects on the planet have been increasingly referred to as a *climate crisis* in lieu of *climate change*. The climate activist group Elokapina – Extinction Rebellion Finland demand the Finnish government to announce a climate state of emergency and to frame the discussion about climate change as a crisis, similarly to COVID-19. The group's aim here is to create a crisis through discourse. That is to say, the existence of climate change alone is not necessarily a crisis. It only manifests as a crisis if it is talked into being one. All this is not meant to suggest that crises are only socially and linguistically constructed, with no relationship to the material world. COVID-19 would still be a deadly disease, infecting and killing people around the world, even if no-one talked about it and climate change will happen even if it is ignored. Rarely, if ever, are crises created completely separate from actual events taking place in the world.

Just as crises disrupt the normal state-of-being, they also disrupt the narratives of that state. Suddenly, the old stories that people have come to rely on no longer provide an explanation for the events taking place. Therefore, a new narrative is needed, to construct a picture a reality that fits the transformed circumstances. By creating and convincing others of the legitimacy of one story and perspective, other, contending voices can be silenced (Heath, 1994). Therefore, whoever manages to tell the most convincing story, holds a lot of power in the way the crisis unfolds and its aftermath. In the age of social media, these stories tend not to be told by single governmental authority figures but rather co-constructed in interaction with other people sharing a social media platform, as noted by Siapera in their case study of Twitter messages containing the hashtag *Palestine* (Siapera, 2014).

The theoretical and empirical understanding of the ways in which narratives on social media emerge and how they differ or are similar to narratives in more traditional communicative settings is still lacking, and there is much research to be done in this relatively new branch of narrative inquiry. Research indicates that small stories are a particularly common form of narrative storytelling in social media posts compared to fully-formed Labovian narratives

(Dayter, 2015). In a case study of social media narratives, Georgalou (2015) employs the small story framework. She analyses Facebook status updates written by a Greek woman, Helen, during the Greek financial crisis in the early 2010s. Throughout the crisis, Helen wrote about political matters, such as elections, her own political activism, and her feelings. Her stories were partly constructed collaboratively with people commenting on her posts and by links to news articles. She also used multimodal communication by combining text with pictures, internet memes and music. Georgalou argues that by sharing their anguish and experiences on social media, raising awareness, evaluating, and responding to the crisis around them, people can position themselves within it. She concludes that "Helen's networked small stories on the crisis give rise to a hybrid subgenre which blends autobiography with news reporting, fact with opinion, subjectivity with objectivity, and emotion with meaning" (p.12).

A narrative view on crisis is particularly relevant because of the power of narratives to mould one's world view and, in effect, one's political view. If narratives are a way to organize our thoughts and understand our experiences, they have immense effect on how we act, re-act and treat the events and people around us. Small, personal stories may have a stronger impact than un-narrativized statistics. As the saying goes, the plural of anecdote is not data. Yet, a well-told anecdote can change minds more effectively than any amount of data. Herein lies the power of small stories to create larger narratives of crisis. Things happening directly to us, impact us more than things happening to others. Conflicts in neighbouring countries garner more sympathy than a war far away. Even though they are often overlooked, small stories matter and researching them in the age of social media is more crucial than ever.

2.3 Digital Discourse Analysis and Corpus-Assisted Discourse Studies

The systematic study of digital discourse is no longer a new field. It started in the 1990s, in the wake of the internet and related technologies such as personal computers and mobile phones becoming more and more commonplace (Herring, 1996). The field has traditionally been dubbed Computer-Mediated Communication (Herring, 1996), although in an age where digital communication takes place more on mobile phones than traditional desktop computers, this term is beginning to show its age. Internet Linguistics is another, broader, term, proposed to cover all linguistic inquiry into internet mediated language (Crystal, 2011). I opt to use the term Digital Discourse Analysis (DDA), because it is not bound to a particular technology, yet places itself firmly in the tradition of the broader field of discourse analysis.

A key issue in studying digital discourse is the rapidly changing digital environments and the wide range of social practices associated with them (Jones et al., 2015). Even though the field itself is not new, it is in constant flux, constantly reinventing itself, as new technologies and discourse domains emerge. However, some researchers also warn against over-enthusiasm for new media and corporate hype (Thurlow & Mroczek, 2011). To tackle the problem of an ever-changing subject of interest, digital discourse analysts have employed a great number of methods and worked within various theoretical frameworks "to address the unique combinations of affordances and constraints introduced by digital media" (Jones et al., 2015, p. 1). Research has been done on a multitude of different internet platforms such as YouTube (Benson, 2015), blogs (Walton & Jaffe, 2011) and comment sections on news websites (Pattamawan & Todd, 2013).

Despite the seemingly boundless breadth of the field, there are some basic tenets that most research in the field abides by. A key concept for understanding DDA is mediation. Scollon (2001) explains mediation to mean the material and social means though which social actors communicate. "A mediated action is carried out through material objects in the world" (p. 4), such as vocal cords or keyboards, clothing, place, and technology. Mediated discourse is also always situated in a unique moment in time and space and can only be understood within a shared understanding of social and cultural practices. All these features of mediation affect how meaning is communicated from one social actor to another. In the context of Twitter, then, we need to understand that technological limitations, social practices, terms of service, time, place, and intended audience all impact how discourse if formed.

An ongoing issue in DDA is its Anglocentrism (Thurlow & Mroczek, 2011). Most research is done in English, using English language data with methods and theoretical frameworks which may not be applicable to other languages and socio-cultural settings. Assuming that American Twitter discourse held in English is identical or even similar to Finnish Twitter discourse is highly problematic. Although efforts have been made to make the field more inclusive of languages and cultures beyond the Anglophone world (Danet & Herring, 2007), much work is still to be done. This thesis is one step in the right direction in this regard, as its subject of analysis is the Finnish language and Finnish socio-cultural sphere.

Corpus-Assisted Discourse Studies (CADS), sometimes known as Corpus-Assisted Discourse Analysis (CADA), and DDA are often overlapping fields of study. Corpus linguistics is a methodology that uses computers find patterns in large bodies of text (or *corpora*, singular *corpus*) (McEnery & Baker, 2015). CADS is a subset of corpus linguistics. CADS combines corpora and quantitative methods developed by corpus linguists with qualitative approaches to study discourse. CADS studies can analyse corpora of any type of language, but of special interest here are studies done on digital discourse, i.e., language-in-use in digital environments, such as social media and messaging platforms. CADS methodologies have been employed for a range of different corpora of digital language. Knight (2015) compared the level of formality in digital language to general language in English, by contrasting the frequency of modal verbs. King (2015) has studied digital sex talk practices using a corpus constructed of chatroom discourse. Lehti et al. (2020) used topic modelling to examine discourses on poverty on the popular Finnish internet discussion forum Suomi24. Twitter is an oft-used source of digital language data. For example, Johansson et al. (2018) studied stancetaking in Twitter messages in the wake of the terrorist attack on the offices of the magazine Charlie Hebdo and the popular hashtag #jesuisCharlie.

2.4 Topic Modelling

Topic modelling is a computational method used to analyse large scale, unstructured data by representing it as a set of topics (Jacobs & Tschötschel, 2019). Over the past decade, topic modelling has been the subject of increasing interest; the number of annual topic modelling publications rose from 75 in 2012 to 292 in 2017 (Li & Lei, 2021). Topic models have demonstrated their utility in content analysis in many fields, such as sociology (R. Huang, 2019), history (Yang et al., 2011) and linguistics (Jacobs & Tschötschel, 2019). Many different topic modelling methods exist, such as Structural Topic Modelling (Roberts et al., 2013) and Dynamic Topic Modelling (Blei & Lafferty, 2006), each with their own specific purposes, advantages and drawbacks.

Latent Dirichlet Allocation (LDA), first described by Blei et al. (2003), is one of the most well-known and most-used topic models (Li & Lei, 2021). LDA functions on a bag-of-words (BoW) type approach. This means that the model does not know the order of the words in a document (in the case of this study, a single tweet), only their frequency. For example, a document which reads ‘Jane kicks a boll’ and another document which reads ‘A boll kicks Jane’ are identical for a model with a BoW approach. LDA assumes that each word in a document is generated by a topic. LDA assigns each word in a text a probability for each topic. Words assigned with the highest probability of a topic are seen as representing that topic (Jacobs & Tschötschel, 2019). Words that co-occur often are likely to belong to one topic, whereas words that rarely occur together are not. One document may, and often does, consist of a mixture of topics. Documents can be ranked by how much of them is made up of each topic.

For instance, if words such as ‘traffic’, ‘pollution’, ‘CO_2’ and ‘climate’ were to co-occur often in a corpus of news articles, they would likely belong to one topic, which could be labelled ‘climate change’. I say ‘could be labelled’ because there is no one correct way to label the topics that the model finds. The model itself only outputs a list of the most common words that belong to a single topic. Determining what the topic is, is left to be done by a human annotator. The ‘climate change’ topic could be found in many news articles, in varying proportions, and one news article could be made up of many topics. An article about melting glaciers would likely be mostly made up of this topic, whereas an article about housing policy would not.

Despite the ever-increasing popularity of topic models as a method of text analysis, some are more critical of the efficacy of topic modelling for the purpose of discourse analysis; Brookes and McEnery (2018) criticize the sole use of word lists in interpreting the topics produced by a topic model and the thematic coherence of the topics. Their study, however, contains a number of methodological errors, mainly, failing to evaluate and optimize their model to fit the data. In the following section, I introduce a number of studies that successfully implement topic modelling to analyse Twitter data concerning the COVID-19 pandemic.

2.5 Previous research on COVID-19 in Twitter discourse

There are a number of studies on internet discourse during the COVID-19 crisis, some of which employ a topic modelling approach. Mutanga & Abayomi (2020) used LDA topic modelling to examine topics discussed among the South African populace on Twitter during March and April 2020. They also monitored local news media and found that similar topics were discusses both on Twitter and in mainstream media. Topics discussed included lockdowns, essential workers, daily infection statistics and conspiracy theories. Janmohamed et al. (2020) used Structural Topic Modelling (STM) and word clouds to analyse how web-based vaping discourse changed after the start of the pandemic. They gathered textual data from of variety of web sources using textual queries related to vaping and e-cigarettes. STM captured major changes in vaping discourse that took place during the time period studied.

The study found that the vaping discourse contained false information about the health benefits of vaping in relation to COVID-19.

In Xue et al. (2020), more than 4 million English tweets collected in March and April 2020 were analysed using LDA and sentiment analysis. 5 themes emerged from the data: measures taken to slow the spread of the virus, social stigma associated with the disease, new cases and deaths, the pandemic in the USA and the pandemic in the rest of the world. Sentiment analysis was used to classify the sentiment or emotion in each tweet. The most prevalent sentiments were anticipation, fear, trust and anger.

Shahi et al. (2021) examined COVID-19 related tweets from a misinformation perspective. They collected fact-checked news articles published between April and July 2020 and crawled through their content to find links to tweets containing misinformation. 1,500 unique tweets, which were categorized as either false or partially false, were found. These tweets were compared to a much larger, general corpus of COVID-19 tweets. They found that false tweet tended to get more likes and spread faster than partially false tweets. Compared to the general corpus, tweet containing misinformation were more likely to discredit information spreading on social media, mentioned governing bodies related to health more often and were more concerned with mortality rates and the latest infection statistics.

Wicke and Bolognesi (2020) collected a corpus of 203,756 tweets over a two week period which contained hashtags relating to the COVID-19 pandemic, such as #covid19, #coronavirus and #ncov2019. They analysed the data using LDA topic modelling to find topics that correlated with framing the pandemic as a war. The WAR framing was most often used in topics of treating and diagnosing the disease. The researchers also compared the prevalence of the frame WAR to other figurative frames, namely STORM, MONSTER and TSUNAMI. WAR was clearly the most prevalent figurative frame but not as prevalent as the literal frame FAMILY. Finally, as new data became available, the research was replicated on datasets that covered tweets gathered over a two-month period. The results were found to be similar to the original research.

A study using Finnish Twitter data discussing the pandemic was conducted by Väliverronen et al. (2020). Their data set consisted of Tweets that mentioned the Finnish Institute for Health and Welfare (‘Terveyden ja hyvinvoinnin laitos, THL’), published between January and June 2020. After a filtration process, 48,459 tweets were queried with keywords and the resulting tweets individually analysed based on a previously established model of alternative expertise on social media. Economic liberalist and data-solutionist critiques of THL's response to the pandemic were analysed and contrasted. The liberalists were found to value lay knowledge and crowdsourcing solutions, whereas data-solutionist called for more open data and employing the expertise of people beyond the field of epidemiology.

The strengths of the current study are multi-faceted. First, few studies have been conducted using Finnish Twitter data. Second, the data in this study cover more than a year of discourse during the pandemic. Most previous studies use data gathered during the start of the pandemic and data that only covers a few months of tweets. The data in this study allows for a dynamic temporal approach that is not present in most previous studies. Lastly, an analysis of crisis narratives has not yet been conducted on similar data.

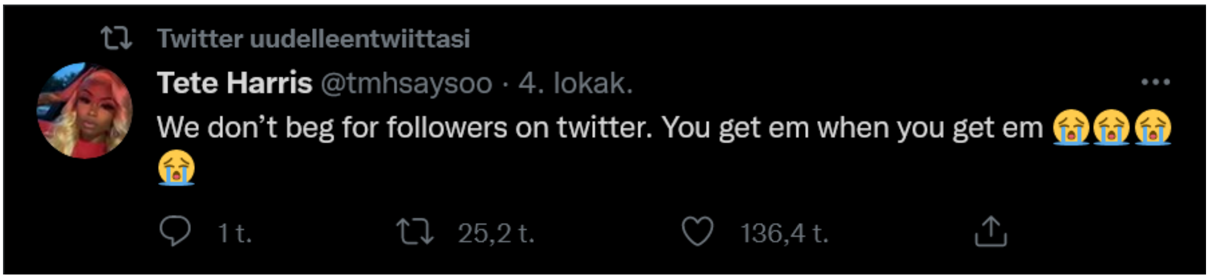
2.6 Some linguistic characteristics of Twitter

To conclude this section on background and theories, a brief discussion of the characteristics of language on Twitter is in order. Just as language used in any social space, be it physical or digital, language on Twitter has its own, distinctive characteristics, affected by the space's social norms, users, and physical and digital limitations. Perhaps the most obvious and distinctive characteristic of language on Twitter is its strict limitation on message length. Originally limiting messages to only 140 characters, later doubling possible message length to 280 characters, Twitter forces its users to write brief messages, often no more than one or two sentences long. This limitation is sometimes circumvented by writing multi-tweet threads.

In addition to length, Zappavigna (2013) notes three distinctive linguistic markers in tweets. These are the use of mentions (@) to address and reference other users, retweeting, and labelling topics with hashtags (#). The @ character is mainly be used to either directly address another user or to indirectly refer to them, for example:

> Direct addressivity: @user I had such a great time last night!
> Indirect reference: Had a great time with @user last night!

Retweets are a way for users to republish tweets written by other users, "effectively recommending it to their followers" (Zappavigna, 2013, p. 43). Retweets were previously marked by the initialism RT, often followed by a reference (@user) to the original author (Zappavigna, 2013). This feature has later been changed by replacing the initialism RT with the text "[user] retweeted" followed by the original tweet, the username of its author, their profile picture, and a link to their account. Figure 3 shows a screenshot of a modern retweet.



Twitter uudelleentwiittasi

Tete Harris @tmhsaysoo · 4. lokak.

We don't beg for followers on twitter. You get em when you get em 😭😭😭😭

Figure 3 Screen capture of a retweet.

In April 2015, Twitter added a new feature called a quote tweet or quote retweet, which "allows users to quote a tweet while adding their own comment" (Garimella et al., 2016, p. 200). Whereas retweets are mainly used to endorse and recommend a tweet, quote tweets can be used to express disagreement (Garimella et al., 2016).

The use of hashtags as a way to label topics possibly emerged from Internet Relay Chat (IRC) and has since become commonplace on many social media platforms (Zappavigna, 2017). Hashtags are a form of user-generated metadata that makes tweets searchable, so that other users interested in the same topic can more easily find them (Zappavigna, 2013). However, users have created innovative ways of using hashtags that go beyond the original use of topic labelling, such as interpersonal communication, showing attitude, and humour (Zappavigna, 2015).

Herring (2007) distinguishes between synchronous and asynchronous communication. In synchronous communication, both the sender and the addressee of a message have to be present or online simultaneously and can hence respond to each other in real time. In asynchronous communication, on the other hand, a message can be sent without the addressee being present to receive it immediately. Twitter communication lands in this latter category. Twitter is also a one-way communication channel, since the addressee can only read and react to a message after it has been fully formulated and sent (Herring, 2007). Compare this to face-to-face communication, in which the addressee can react or interrupt mid-sentence. However, communication of Twitter can be very fast paced, since users with mobile phones can receive notifications about new messages and respond to them immediately.

Another dimension to categorize computer-mediated communication is by the number of intended recipients (Baron, 2008). One-to-one communication is private communication between to people. Examples of such are emails and text messages. One-to-many communication entail messages meant to be read by a larger audience. Tweets are examples of one-to-many communication because they are publicly available to anyone. Although, as discussed above, the @ character can be used to directly address a single user, the message is still readable by other users.

3 Material and Methods

In this section, I explain the data gathering process, how the data were manipulated and cleaned, and how the topic modelling was performed. First, I describe how and what data was gathered and clarify the steps in data pre-processing and justify the reasoning behind the decisions that were made. I also give descriptive statistics of the data, which already illustrates broadly how the discussion around COVID-19 evolved over time.

3.1 Ethical considerations

The data used in this study come from public messages sent on Twitter between January 2020 and August 2021. The tweets are publicly available on the Internet and were gathered using the Twitter API for research purposes. Even though the tweets analysed in this study are publicly available, user information is not disclosed, except when this information is important for a full understanding of the context of the tweet or the user is a public figure or particularly active tweeter, in accordance with the guidelines of the Finnish National Board on Research Integrity (2019).

3.2 Material

3.2.1 Twitter

Twitter is a micro-blogging service and digital environment that allows users to write short messages (known as tweets) up to 280 characters long. These messages can be seen, reacted to, commented on and shared by other users on the platform. Users can choose to follow accounts that they find interesting and thereby have their tweets appear on their feed. Twitter is one of the most popular social media sites both globally and in Finland. In the second quarter of 2020, Twitter claimed 186 million monetizable daily active users (Twitter, 2020). In Finland, Twitter is the fifth most used social media service, with 13 percent of the population using it (Official Statistic of Finland, 2020). Ahead of Twitter in popularity are Facebook, the most popular social media service in Finland, reaching 58 percent of the population, followed by WhatsApp (53%), Instagram (39%), and Snapchat (14%). Young people are more likely to be Twitter users, with 27% of 16- to 24-year-olds reporting to use the service (Official Statistic of Finland, 2020). The older the person, the less likely they are to use Twitter: 20% of 25- to 34-year-olds, 18% of 35- to 44-year-olds and 14% of 45- to 54-

year-olds use Twitter. Usage drops even lower in the older population. Men are slightly more likely to use Twitter than women, with 16% and 11% of them using the service, respectively.

Although Twitter usage in Finland is low compared to Facebook and WhatsApp, Twitter has become an important platform for socio-political discussion (Ojala et al., 2019). The platform has been characterised as a network of political elites, consisting of politicians and media personnel who follow each other and whose tweets gain lots of attention (Vainikka & Huhtamäki, 2015). Twitter is also a popular platform for so-called hashtag politics and slacktivism. Most famously, although the catchphrase MeToo was coined a decade earlier, in 2017 #MeToo entered mainstream discourse around the world after actress Alyssa Milano used it in a tweet encouraging victims of sexual harassment to speak up (Schneider & Carpenter, 2020). Hence, even though Twitter is not the largest social media platform, it is a hub for political discourse and societal critique. This, combined with the relative ease of access to content on the service, makes Twitter particularly attractive platform for sociolinguistic analysis.

3.2.2 Data gathering

The data used in this research were gathered using Twitter's Application Programming Interface (API). The API allows for real-time gathering of tweets that match user given search terms. Data gathering started on April 1st 2020 using a search term list, which included words that were related to the COVID-19 pandemic, its most typical symptoms, and other topics that were discussed in the media and society at large at the start of the pandemic (see Appendix 1 for complete list). Additionally, tweets that were sent earlier than April 1st were later searched for using the same search term list and added to the dataset. Data gathering was stopped on August 6th 2021.

The dataset was originally gathered for a different implementation than what it is used for here. The original use purpose was to follow the development of the pandemic in near real-time by tracking the location and prevalence of people tweeting about suffering from typical symptoms of COVID-19. Therefore, the search term list includes terms such as *hiki* ("sweat"), *päänsärky* ("headache") and *ripuli* ("diarrhoea"). The list also has similar search terms in Swedish. Only Finnish tweets are analysed in this study.

3.2.3 Data pre-processing

The tweets that were gathered using Twitter's API were in JSON format. The raw JSON files were processed to access relevant information and discard information that was not relevant for this study. This was done using a Python package called TweetParser[1]. This software allowed for convenient extraction of relevant metadata and user-entered text.

The dataset originally consisted of 1,339,416 tweets. Out of these, 5,837 were sent out in the year 2019. Although the virus began spreading in China in 2019, it only became a major subject of discourse in Europe in 2020. Therefore, this study only examines tweets posted between January 1st 2020 and August 6th 2021. Retweets were also removed from the data, which reduced the dataset down to 685,376 tweets. Retweets were removed, since they would add unnecessary duplicates to the data, which could skew the results. Quote tweets were included but the quoted parts were removed and only user entered text was kept. Then, duplicate tweets were removed. By duplicates, I refer to tweets that had the same tweet ID number. These are tweets that were erroneously added to the dataset more than once. Tweets with identical content but different ID numbers were kept.

As the dataset was originally collected as a part of a project to examine the real-time spread of COVID-19 in Finland, the data were tailored for the requirements of the current study. A new list of search terms was created in order to exclude as many irrelevant tweets as possible and only to include tweets that overtly refer to the pandemic and measures to reduce its spread. A cursory exploration of the dataset set was conducted to find words that were often used in tweets about COVID-19. Additionally, words relating to topics that were often discussed in the media during the pandemic were considered. Eventually, a purposefully short list of search terms was created. The list was kept short because the original dataset contained many irrelevant tweets. Only tweets which included one or more of the following search terms were included in the final dataset: *korona, corona, covid, karanteeni, maski, tartunta, rokote, rokotus, pandemia, epidemia*. The final dataset consisted of 375,322 tweets.

Before using the data as input for a topic model, further pre-processing was needed. The tweets were lemmatized and parsed using the Turku Neural Parser Pipeline (Kanerva et al., 2018). Lemmatization returns every word to its 'basic' form, i.e., lemma. For instance, *writes*, *writing*, and *wrote* are all inflected forms of the same lemma, *write*. For the purpose of topic

[1] pypi.org/project/tweet-parser/

modelling, it is better to treat these inflected forms as instances of one word rather than three different words, which is what lemmatization achieves. Because the lemmatization was done computationally, the parser produced some misslemmatizations. The lemmas 'korona', 'korona#virus', and 'karanteeni' (*quarantine*), being rarely used words in common discourse before the pandemic, were often wrongly lemmatized as 'korkona', korko#avirus' and 'karanneeni', respectively. However, since this misslemmatization was quite systematic, its effects on the quality of the topic model are minor. After the data were lemmatized, the following steps were taken to further clean the data:

1. Lemmas were normalized into lowercase.
2. Hyperlinks and single character words were removed.
3. Stop-words were removed. Stop-words are words that occur very often but tell very little of the meaning of a document, such as *and*, *but*, *in*, etc. Here, conjunctions, adverbs, adpositions, symbols and punctuation were regarded as stop-words.
4. Words that occurred in only one document were removed as were words that occurred only two time or fewer in the entire corpus.

Figure 4 below visualises the process of removing irrelevant data and pre-processing the final dataset.

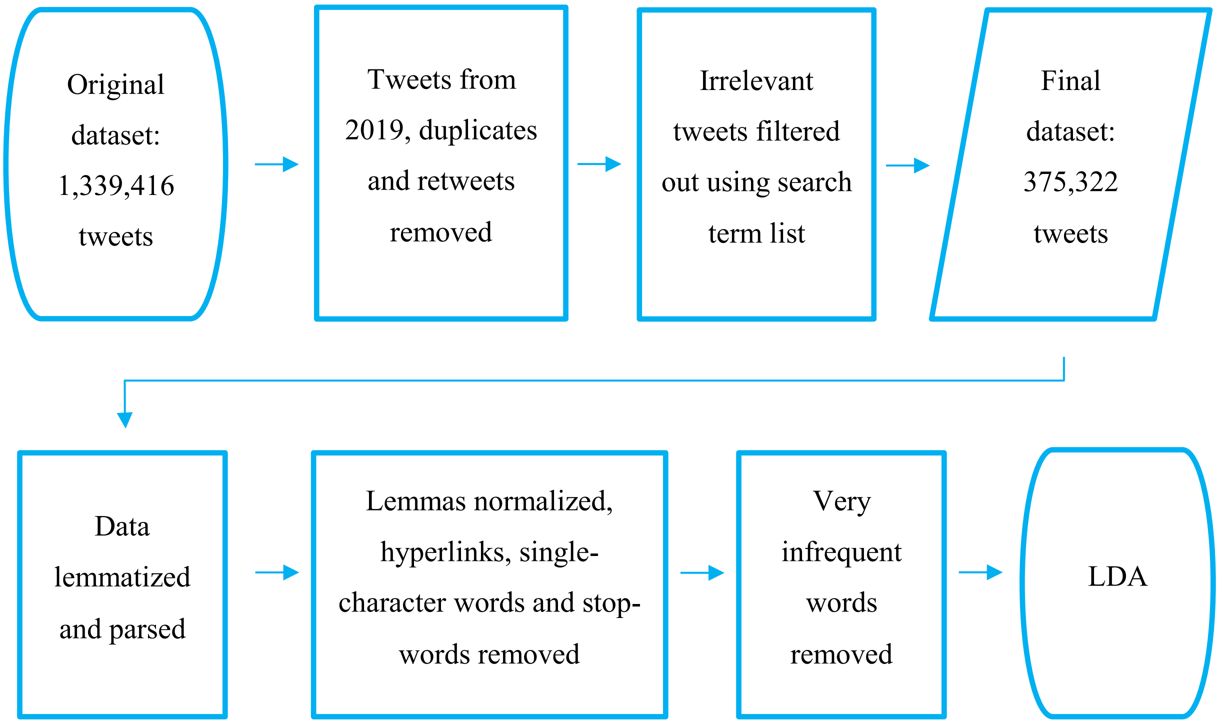

Original dataset: 1,339,416 tweets
Tweets from 2019, duplicates and retweets removed
Irrelevant tweets filtered out using search term list
Final dataset: 375,322 tweets
Data lemmatized and parsed
Lemmas normalized, hyperlinks, single-character words and stop-words removed
Very infrequent words removed
LDA

Figure 4. Data cleaning and pre-processing

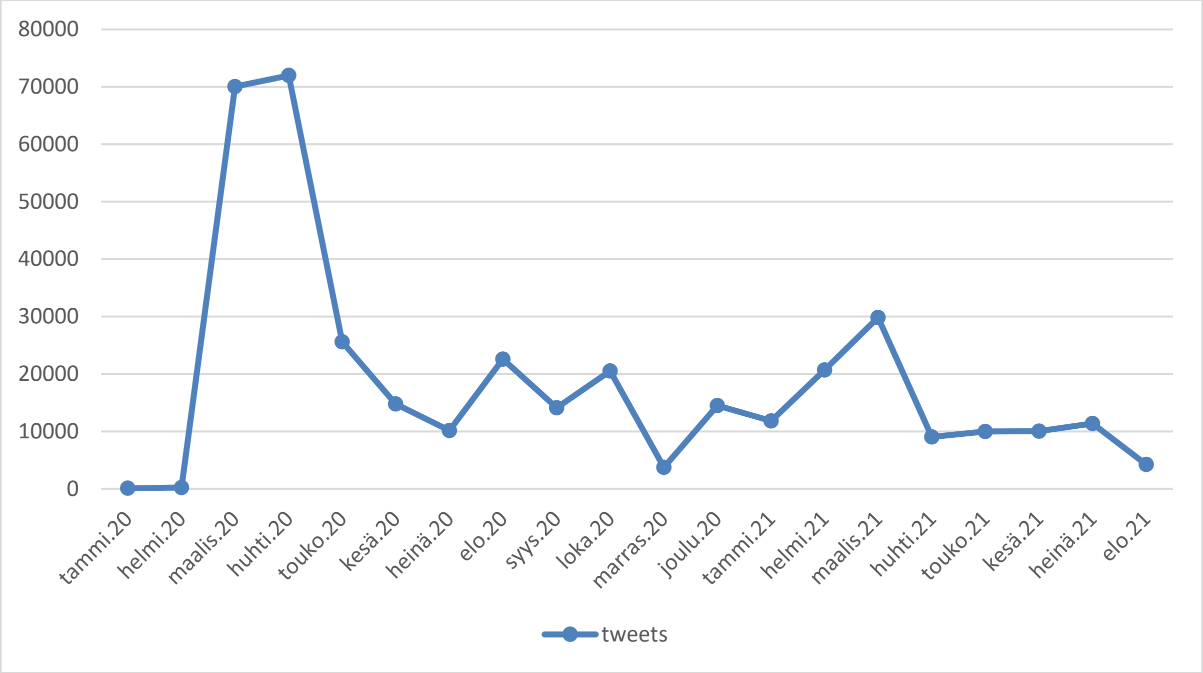
3.2.4 Descriptive statistics of the data

Out of the total 375,322 tweets that were left in the final dataset after the removal of unwanted data, 307,203 were conventional tweets and 68,119 were quote tweets. These tweets consisted of a total of 4,477,614 words with an average length of 11.93 words per tweet. The search term ‘korona’ was by far the most common in the final dataset. It occurred in 70.19% of tweets. The second most common search term, ‘tartunta’ (*infection)* occurred in only 7.64% of tweets.

search term	**tweet (*n*)**	***N***	**%**
korona	263425	375322	70.19
tartunta	28684	375322	7.64
maski	23516	375322	6.27
covid	20222	375322	5.39
rokote	18193	375322	4.85
karanteeni	17079	375322	4.55
epidemia	12388	375322	3.30
pandemia	11834	375322	3.15
rokotus	10399	375322	2.77
corona	4342	375322	1.16

Table 1. Document frequency of search terms in final corpus.

The distribution of tweets by month is shown below in Figure 5. The chart shows very clearly that there was very little discussion around COVID-19 in January and February of 2020, only 81 and 198 tweets in this data set, respectively. Using Fink’s (1986) taxonomy, these two months represent the prodromal stage of crisis. In March 2020 the number of tweets leaped to 70,049 tweets followed by 71,970 tweets in April 2020. These two months of high activity represent the acute crisis stage. After this, monthly tweets dropped for the remainder of the data collection period, not surpassing 30,000 monthly tweets. This long period can be seen as the chronic crisis stage. The resolution stage of crisis was not reached during the time period when the data were collected. Due to the evolving nature of crisis, the data are not spread equally temporally. In fact, almost 38% of all tweets in the data were sent during March and April 2020.



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2020	81	198	70049	71970	25577	14782	10148	22589	14118	20508	3751	14508
2021	11825	20692	29832	9044	9984	10042	11371	4253	N/A	N/A	N/A	N/A

Figure 5. Number of tweets by month.

3.3 Method

The methodology of this study is a combination of both quantitative and qualitative methods. First, the full data set is quantitatively analysed by means of topic modelling, more specifically Latent Dirichlet Allocation. Based on the results of the quantitative method, a thorough close-reading of tweets that correspond to the most prominent topics is undertaken. Employing a theoretical framework of narrative studies, literature on crisis and digital discourse analysis, the narrativization of crisis and its evolution throughout the first 20 months of the COVID-19 pandemic is examined.

3.3.1 Topic modelling and model evaluation

The LDA topic modelling was performed using the Gensim Python library (Rehurek & Sojka, 2010). The computation was done on the supercomputers Mahti and Puhti, provided by the IT Center for Science[2]. After initial test runs, a pipeline was created, in which the data were pre-processed and provided as input in the correct format for the LDA model. The input was

[2] https://www.csc.fi/en/home

vectorised and transformed into a bag-or-word representation. The model was run with 10 passes and a chunksize of 2000. Passes refer to the number of times the entire training corpus is passed through the model and chunksize refers to the number of documents (i.e., tweets) the model processes at a time. The alpha and eta values are the assumed document-topic and topic-word distributions in the corpus, respectively. The optimal values for these hyperparameters were set to be automatically derived from the data.

A crucial step in initializing a topic model is deciding the total number of topics (*K*). This step is somewhat arbitrary and has to be determined on a case-by-case basis. The appropriate value of *K* depends on the size of the data set and the contents of the documents within it. An appropriate value for *K* can be found using a trial-and-error method, i.e., running the model multiple times with different values of *K* and analysing the results produced by each run. If the topics seem too vague and general, there are likely too few of them. If, on the other hand, the topics seem hyper specific and hard to distinguish from one another, there are too many topics. This kind of approach can produce very human-interpretable topics, but at the cost of either being highly subjective, if done by only one person, or very costly, if done by a large group of human evaluators (Röder et al., 2015).

Researchers have addressed this issue by developing a multitude of statistical methods for (semi-)automatic model evaluation (see Arun et al., 2010; Nikolenko et al., 2017; Röder et al., 2015). These methods have promised to make model evaluation faster and more objective. Initially, perplexity-based measures were used for evaluation. However, these measures are now disfavoured because they were found not to correlate with human evaluators (Chang et al., 2009). That is to say, topics that were rated well by the perplexity measures, were found to be somewhat incoherent by humans. Röder et al. (2015) have developed a new coherence metric C_V, which is a combination of previously developed coherence measures. In their study, C_V performed better than other tested methods in correlating with human evaluations of topic coherence. This is the measure used in this study to evaluate model performance. To find the optimal number of topics, a series of models with different numbers of topics were run. Models were run at 10 topic intervals to find the best fitting model. The C_V coherence score for each model was calculated and the best fitting model was found to be a model with 110 topics. Figure 6 visualises the coherence score measured for each value of *K*.

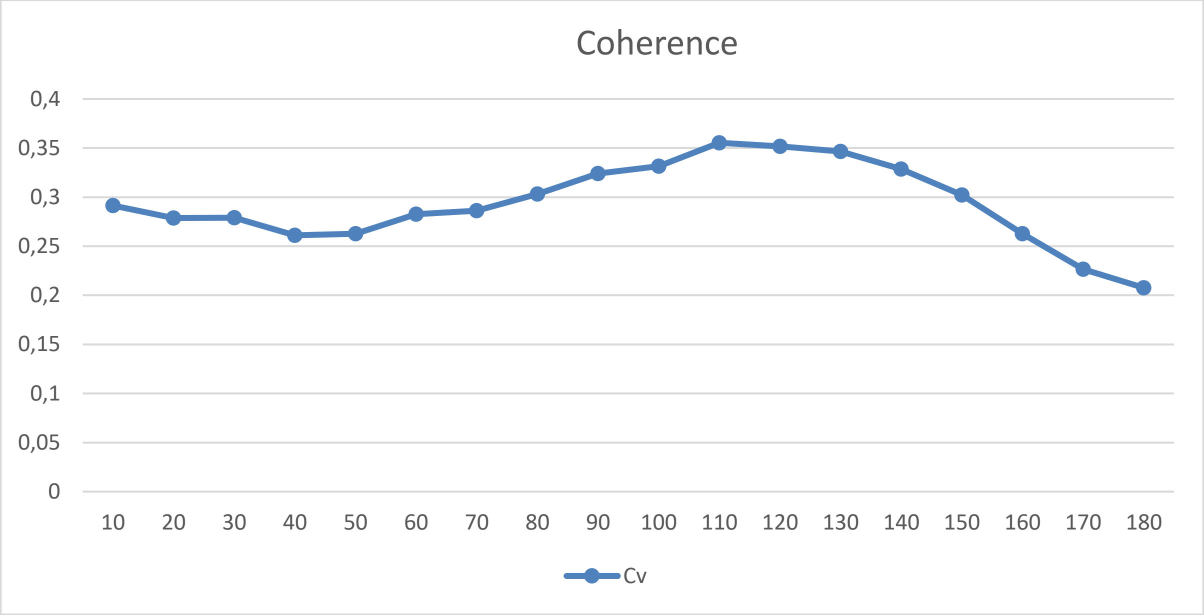


Figure 6. C_V scores for model run with different values of *K*.

Topic coherence measures are a useful tool for optimizing a topic model. However, there are a many coherence measures that give varying results on different types of data. Topic coherence measures are not to be taken as the absolute truth, but only to be used as guides. It is the responsibility of the researcher to determine how well the chosen measure correlates to topic quality by closely examining the topics the model proposes. Therefore, after finding the model with the best coherence score, the topics it produced were manually confirmed to be coherent. This was done by analysing 25 keywords for each topic. Keywords are the most relevant words for a given topic, i.e. they are assigned the highest probability for that topic. Topics were considered coherent when the keywords together formed a meaningful whole with only few outlier words that could not easily be semantically linked to other keywords.

3.3.2 Finding most frequent topics

The output of the topic model was 110 topics, each represented by 25 keywords. To examine the distribution of topics in the data, first the distribution of topics in each tweet was calculated using the Gensim library. As explained above in section 2.4, each tweet consists of a distribution of topics. Therefore, it is possible to calculate which proportion of a tweet is covered by a specific topic. For each tweet, the five most prevalent topics were chosen to represent the contents of the tweet. Then, the sum of the number of times each topic was among these top five topics was calculated. To explain this more concretely, let’s imagine a dataset of only three tweets: A, B and C. If topic *a* was the most prevalent topic in tweet A, the third most prevalent topic in tweet B but only the sixth most prevalent in tweet C, the sum of topic *a*’s appearances among the top five topics would be 2. Using this method, the topics that were most relevant in the data were found. Figure 7 depicts this method used on a hypothetical corpus. This was done because examining all 110 topics was deemed unpractical and unattainable for the scope of this study.

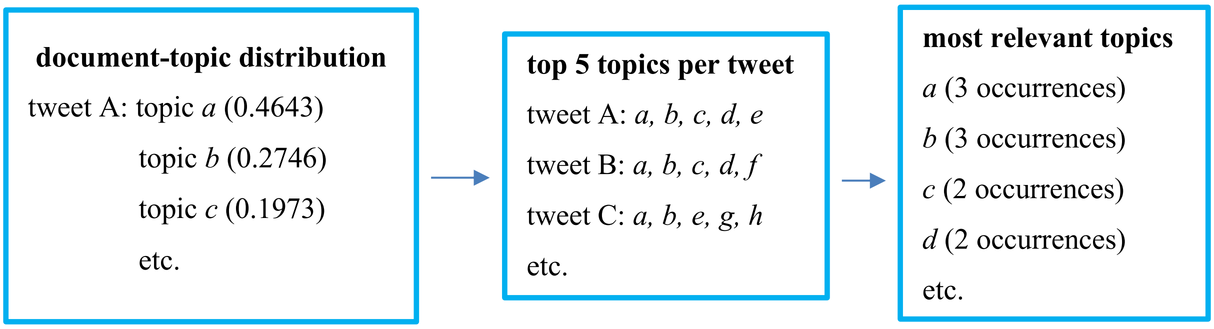


Figure 7. Finding most relevant topics

The 21 most frequent topics were then labelled. Labelling was done by manually examining the keywords for each topic and tweets that were strongly associated with the topic. The strength of the association was determined by using the topic-document distribution of a given document. These labelled topics are listed below along with the number and proportion of tweets in which they were among the top five. I chose to analyse the 21 most frequent topics because topic 80 was found to be in the top five of every tweet in every time slice. This is because the topic consisted of very general COVID-19 related discussion, which is what the data were constructed to represent. The keywords for topic 80 included words such as ‘olla’ (be), ‘ei’ (no), ‘korona’, ‘se’ (it), ‘koronafi’, ‘tämä’ (this), ‘voida’ (be able to), ‘mikä’ (what), ‘ihminen’ (human), and ‘korko#avirus’ (misslemmatization of ‘korona#virus’). Topic 80 tells more about the general nature of the data and little about the particular nature individual tweets. Hence, it will be largely ignored. The existence of this topic does, however, lend credibility to the validity of the topic model. It shows that the model has learned to identify what the unifying theme of the corpus is. Therefore, it is reasonable to assume that the other topics suggested by the model also accurately represent the contents of the corpus.

3.3.3 Organizing data temporally

To analyse how the latent topics found using LDA change in popularity as the crisis evolved, the data were arranged into sequential order and split into time series. These splits are hence referred to as time slices. Due to the uneven distribution of the data (see Figure 5 in section

3.2.4), a simple split on a monthly basis was not possible, because the number of tweets in each time slice would have varied greatly from as much as over 70,000 to as few as less than a hundred. To address this problem, the two months that had much more data than the others, March and April of 2020, were designated to their own time slices, whereas all other months were combined into two-month blocks. Despite this grouping of months, the time slices are not equal in size. However, making the time slices larger would mean losing detail in the changes in distribution of topics over time. As should be clear from the overview of the development of the COVID-19 pandemic in Finland given in 2.1, at times, the crisis developed very rapidly, and hence, combining more than two months into one time slice would risk obscuring these fast-paced temporal changes in the pandemic.

To contextualise the time period each time slice represents, I assigned each to a stage of crisis, based on the model by Fink (1986). Time slice 1 is stands out for two reasons. Firstly, it is much smaller than any other time slice, with only 279 tweets. This makes direct comparison of topic distribution between this time slice and others unreliable, since only a handful of tweets could potentially result in a large spike in the proportion of a topic. Secondly, it represents the time period before the beginning of the crisis proper (i.e. the prodromal stage). Time slices 2 and 3 are unique since they only cover the time period of one month each. They represent the acute stage of crisis, the beginning and immediate reaction to a radical shift in people's lives. The rest of the time slices represent the chronic stage of crisis. The resolution stage was not reached during the data gathering period.

Time slice number	**Months in time slice**	**Stage of crisis**	**Number of tweets in time slice**	**% of total tweets**
Time slice 1	January and February 2020	Prodromal stage	279	0.07
Time slice 2	March 2020	Acute stage	70,049	18.66
Time slice 3	April 2020	Acute stage	71,970	19.18
Time slice 4	May and June 2020	Chronic stage	40,359	10.75
Time slice 5	July and August 2020	Chronic stage	32,737	8.46

Time slice 6	September and October 2020	Chronic stage	34,626	9.23
Time slice 7	November and December 2020	Chronic stage	18,259	4.86
Time slice 8	January and February 2021	Chronic stage	32,517	8.66
Time slice 9	March and April 2021	Chronic stage	38,876	10.36
Time slice 10	May and June 2021	Chronic stage	20,026	5.34
Time slice 11	July and August 2021	Chronic stage	15,624	4.16

Table 2. The time period and stage of crisis each time slice represent and the number and proportion of tweets in each time slice.

3.3.4 Analysing topics

After the 21 most frequent topics were labelled, the 10 most frequent (ignoring topic 80) of those were chosen to be analysed more thoroughly, because a detailed analysis of all 20 topics proved to be too impractical for this study. The ten most frequent topics were analysed within the theoretical framework of crisis and narrative studies. The narrativization of crisis was analysed both in terms of a broader, overarching narrative structure, as proposed by Hay (1996) and as small stories (Georgakopoulou, 2017). This approach was chosen because small stories often report ‘breaking news’, immediate unfolding events. Twitter prompts its users to share such narratives by asking ‘What’s happening?’ and the Finnish equivalent ‘Mitä tapahtuu?’ at the top of the website. The present tense verb in the prompt encourages users to write stories about the here and now and the strict character limit of a tweet forces users to write short and concise messages. This form of communication fits well under the analytical lens of small stories.

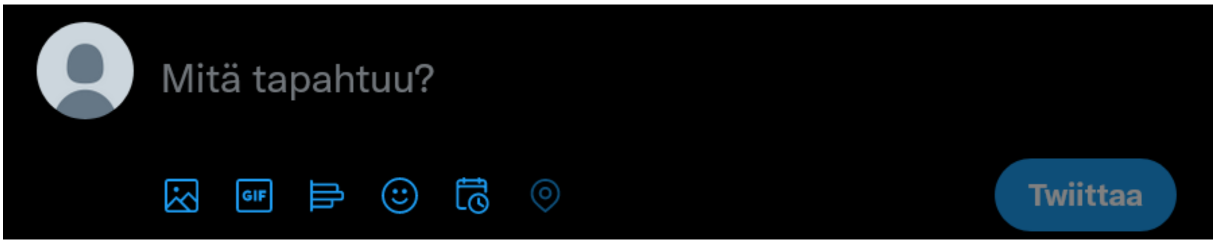


Figure 8. A screen capture showing part of the user-interface of Twitter.

I recognise the need to understand the larger context in which a single narrative is authored. A small story cannot be understood in a vacuum (Georgalou, 2015). It must be read in the contexts in which it was created. The narrative of a single tweet must be read as part of a larger meta-narrative, as suggested by Hay (1996). Each tweet builds upon previous narratives told in other tweets, news stories and beyond, forming a nexus of narratives that together form a whole greater than the sum of its parts. Small stories become large stories through their intertextual connection to other similar stories.

4 Results

In this section, I analyse the data quantitatively and qualitatively. First, I provide frequency measures of the most active tweeters and most common hashtags to get a better understanding of the data and gain some cursory insight into the kind of discourse the data contain. Then, the results of the topic model and the distribution of topics over time are analysed.

4.1 Most active tweeters

In this section, I provide a breakdown of the data over individual tweeters. The more a person tweets, the more involved they are in the discussion and the more likely their tweets are to be seen by a large number of people. As a reminder, the data consist of 372,322 tweets. These tweets were sent from a total of 42,866 unique user accounts. 18,906 or 44.1% of these accounts contributed only a single tweet. 35,263 or 82.26% of users contributed 10 or fewer tweets. The median number of tweets from one account was only two. From these numbers it is clear that a vast majority of users only rarely engaged in COVID-19 related discourse. On the other end of the scale, there were 17 accounts that contributed more than 1,000 tweets each. These accounts are listed below in Table 3. In total, these 17 most active accounts constitute 24,927 tweets or 6.64% of all the data.

Because these active tweeters can be highly influential in the kinds of discourse and the kinds of narratives that emerge on the platform, they are analysed further. To gain insight into the who the most active tweeters were and how they present themselves on Twitter, I visited their Twitter profile pages. Based on these public profiles, the accounts were categorized by user type, such as politician and news media. Accounts that were not linked to public figures and claimed no particular expertise in a relevant field were categorized as private users. Private users are anonymized to protect their privacy. The account N2 seems to have been deleted after the data collection was finished, because it could no longer be found. Therefore, its user type is unknown.

Among the most active tweeters, there were three accounts labelled news media, one labelled political media and one labelled other media. The account Vastuullisuus was labelled other media, because it does not represent a major established news media, but shares links to news articles produced by others. Two of the most active tweeters were bot accounts. Bot accounts sent automated messages that are not written by humans. The account bot_fi tweets daily statistics about COVID-19, such as new infections, deaths, and hospitalisations. FinnaBot tweets pictures based on the most popular hashtags. The two experts on the list were lasleh, the account of Lasse Lehtonen, an expert at Helsinki University Hospital (HUS) and merjah, the account of Merja Helle, retired head of Media Concepts Research Group at Aalto University. The two politicians on the list are not major party leaders or well-known senior politicians.

Username	**User type**	**Number of tweets**
uusisuomi	news media	2,389
j_martikainen	politician	2,185
lasleh	healthcare expert	2,061
Vastuullisuus	other media	1,777
MarkoTJEkqvist	politician	1,540
bot_fi	bot	1,533
N1	private user	1,478
turunsanomat	news media	1,397
N2	unknown	1,291
merjah	other expert	1,267
N3	private user	1,217
N4	private user	1,212
FinnaBot	bot	1,203
suomenmaa	political media	1,198
riikka_kevo	private user	1,096
talouselama	news media	1,054
somevaari	private user	1,029

Table 3. Most active users in dataset.

4.2 Most common hashtags

A query into the most prevalent hashtags in the data was carried out. Since hashtags can be used as topic tags on Twitter (Wikström, 2014), a cursory investigation of the topics discussed in the dataset can be done by simply looking at the most common hashtags. All hashtags were normalized to lowercase and sorted by frequency. In total, there were 904,510 hashtags and 83.9 percent of all tweets contained at least one hashtag. There were 70,731 unique hashtags, most of which were used only once. In fact, 45,656 unique hashtags make just a single appearance in the data.

Not surprisingly, given the search terms used to gather the data, the most used hashtags directly mentioned ‘korona’ or ‘covid’. By far, the most common hashtags were ‘korona’ and ‘koronafi’ which accounted for 12.68% and 10.66 of all hashtags, respectively. To get a more nuanced understanding of the types of hashtags used in the data, hashtags that included the strings ‘korona’ and ‘covid’ were temporarily removed. The results of this analysis show that the 10 most common hashtags were ‘hallitus’ (*government*), ‘talous’ (*economy*), ‘politiikka’ (*politics*), ‘thl’, ‘suomi’ (*finland*), ‘karanteeni’ (*quarantine*), ‘etätyö’ (*remote work*), ‘pysykotona’ (*stayhome*), ‘rokote’ (*vaccine*), and ‘helsinki’.

Most common hashtags (*N* = 904,510)			**Most common hashtags with mentions of ‘korona’ and ‘covid’ removed (*N* = 516,451)**		
Hashtag	*n*	%	Hashtag	*n*	%
korona	114,725	12.68	hallitus	11,051	2.14
koronafi	96,421	10.66	talous	5,741	1.11
koronavirus	49,694	5.49	politiikka	5,718	1.11
covid19	34,193	3.78	thl	4,342	0.84
koronakriisi	29,387	3.25	suomi	4,316	0.84
koronasuomi	18,671	2.06	karanteeni	4,151	0.80
hallitus	11,051	1.22	etätyö	3,581	0.69
covid19fi	9,841	1.09	pysykotona	3,493	0.68
koronavirusfi	7,898	0.87	rokote	3,414	0.66
talous	5,741	0.63	helsinki	3,336	0.65

Table 4. Most common hashtags in dataset.

4.3 Topics

The topics were labelled based upon the top 25 keywords in each topic. Additionally, full tweets that were strongly associated with a topic were examined to gain a better understanding of the topic, if the keywords were not enough to determine a label. Major themes arising from these topics are regulations enforced to halt the spread of the pandemic (topics 50, 26, 22, 83, 77), the effects these regulations had on both people’s personal lives and society at large (32, 24, 99, 94) and discussions regarding the symptoms and effects of the disease (0, 4, 65, 49, 79, 47, 106, 42, 101). Notably, the topics are not equally frequent in the data. Topic 50 is by far the most common, occurring in the top five most frequent topics of nearly 99% of tweets. The second most common topic, topic 26, is far less common (30.89%).

Rank	**Topic label**	**Topic number**	***N***	**%**
1	Restrictions and regulations	50	370913	98.93
2	Quarantine	26	115937	30.89
3	Altercations and foreign countries	32	74794	19.93
4	Border controls	22	54095	14.41
5	Fear of death and sickness	0	46287	12.33
6	Public spaces and services	83	43226	11.52
7	Development and spread of pandemic	64	39833	10.61
8	Testing and symptoms	4	35750	9.53
9	Work life	24	35486	9.45
10	Government briefings	77	32359	8.62
11	Spread of disease	65	29105	7.75
12	THL and healthcare	49	27020	7.20
13	Descriptions and feelings	79	25540	6.80
14	COVID-19 in the media	7	24435	6.51
15	Economy and international news	99	23174	6.17
16	Economy and politics	94	21454	5.72
17	Feeling ill	47	20466	5.45
18	Sickness and age	106	18214	4.85
19	Sickness prevention and childcare	42	17532	4.67

20	Sickness and healthcare	101	17029	4.54

Table 5. 20 most frequent topics.

4.4 Word clouds

I use word clouds to visualize the keywords in topics. In a word cloud, words that occur more often or are deemed more important are displayed in a larger font and more centred than words that occur less or are less important. In the case topic modelling, each topic can be represented by a list of words that each have a probability measure attached to them. The higher the probability, the more central the word is in that topic. Each word cloud below consists of 25 words that had the highest probability of appearing in the corresponding topic. The larger the word appears, the higher the probability measure. The colours are arbitrary and only serve to make the word clouds easier to read. The words are shown in their lemmatized form. The medial # symbol (e.g. in ‘teho#hoito’) indicates that the word is a compound word.

4.5 Topic analysis

In this section, I analyse the 10 most frequent topics in the data, excluding topic 80. The topics are visualised with a graph and a word cloud. The graphs show changes in the percentage of tweets in which a topic was among the top five most frequent topics. For each topic analysed below, I discuss the topic based on its keywords and individually chosen tweets with a comparatively high distribution of the relevant topic. I also discuss the wider context of the topic and suggest possible reasons for the changes in its trajectory over time.

4.5.1 Topic 50: Restrictions and regulations

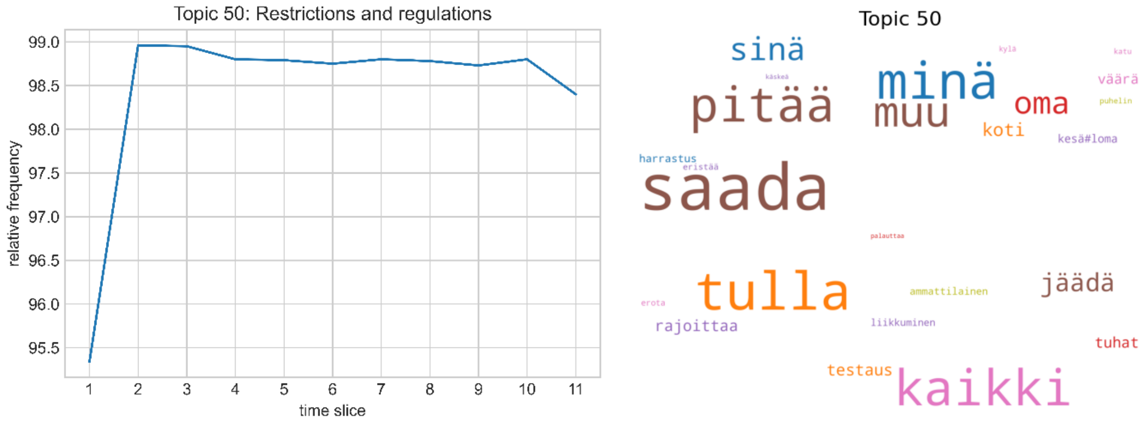


Figure 9. Temporal distribution and keywords for topic 50

Topic 50, ‘Restrictions and regulations’, is a very frequent topic in all time slices. Its prevalence is partly explained by the inclusion of such keywords as the very common first and second person singular pronouns ‘sinä’ and ‘minä’. A reading of tweets that were strongly associated with this topic reveals that users often write about COVID-19 regulations in relation to themselves. In (1), a user informs that they and other inhabitants of the Uusimaa region are or soon will be isolated from the rest of the country, referring to the restrictions placed on travelling to and from Uusimaa in March and April 2020. The rising hands emoji, often signalling joy or excitement, is here used sarcastically.

(1) Minut ja kaikki muut uusimaalaiset on nyt eristetty, tai tullaan pian eristämään. 🙌🏻 #korona

A form of social control was exerted via a criticism of those who did not follow the restrictions and guidelines that were in place. Accusations of self-centeredness were made against those who wanted to return to life as usual. A conflict of narratives is observed in example (2), where a user mocks those, who are believe the crisis is over and by doing so tries to prolong the crisis narrative. Narratively this division of people into good, responsible citizens and bad, selfish and irresponsible citizens is similar to the division between protagonists and antagonists; a narratively convenient and satisfying arrangement that allows for the othering of the ‘bad guys’ and self-aggrandizement of the self.

(2) ”Kaiken keskiössä on se suuri minä, jota poliitikot haluavat miellyttää. Minä olen varannut hiihtoloman Lappiin, joten minun pitää sinne päästä. Minä olen kyllästynyt ravintolarajoituksiin, joten minun pitää päästä yökerhoon. ” Jne., lue👇#koronasuomi #hiihtoloma # ravintolat

The restrictions and regulations put in place to halt the spread of the virus were to many the clearest and most concrete sign of crisis. Whereas, for a long time, the disease itself remained a distant, almost ethereal being, its effects were felt by all. To most, the crisis was made real not by fever, cough, lung damage and death, but by the concrete changes everyone had to make in their daily lives. As discusses above, a crisis is, first and foremost, a sudden deviation from the norm. As this change happened to all simultaneously, stories of how its effects were felt personally were highly relatable and, hence, a popular form of narrative.

4.5.2 Topic 26: Quarantine and vaccination

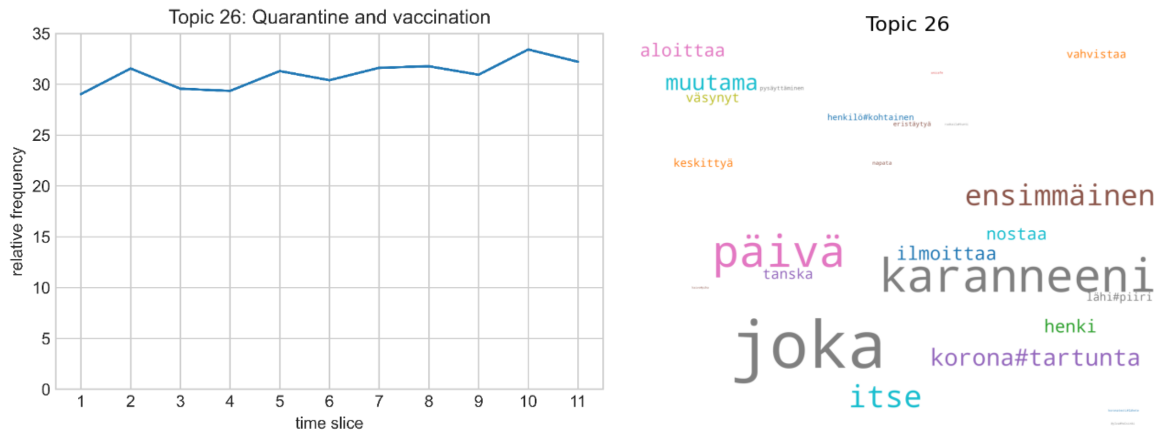


Figure 10. Temporal distribution and keywords for topic 26

Topic 26 centres around discussion about quarantines, vaccinations, and symptoms of the disease. Although the word ‘rokotus’ (*vaccination*) does not appear in the keywords, a reading of tweets that were strongly associated with this topic reveals that the keyword ‘ensimmäinen’ (*first*) often occurs in vaccine related discourses, e.g., ‘ensimmäinen piikki’, (*first shot*). This topic is also very central in the data as a whole, occurring among the top five topics in almost a third of all tweets. As with topic 50, this topic also concerns sudden deviation from how things used to be. Being forced into quarantine, by law or by peer pressure, was a completely novel experience to most.

This topic is on a generally rising trajectory. It first peaks in time slice 2, as the crisis begins, and quarantines are first implemented on a large scale. Another peak occurs in time slice 10. This peak corresponds with much-discussed incident of Finnish football fans travelling to Russia to watch the Finnish team play in the UEFA Euro 2020 tournament. The novel Delta variant had not yet been spreading in Finland but was spreading rapidly in Russia. After the fans returned to Finland, infection rates started growing. In (3), a user discusses quarantining and expresses their belief that most fans would follow the instructions regarding it

> (3) @Huuhkajat Karanteeni se on omaehtoinenkin karanteeni. Varmasti suurin osa porukasta sitä noudattaakin. Kolme päivää karanteenissa ja sitten testiin, josta negatiivisen saatua vapautuu karanteenista.

Additionally, rising infection rates and the beginning of the mass-vaccination campaign in early 2021 potentially led to a rising trajectory in this topic.

4.5.3 Topic 32: Altercations and foreign countries

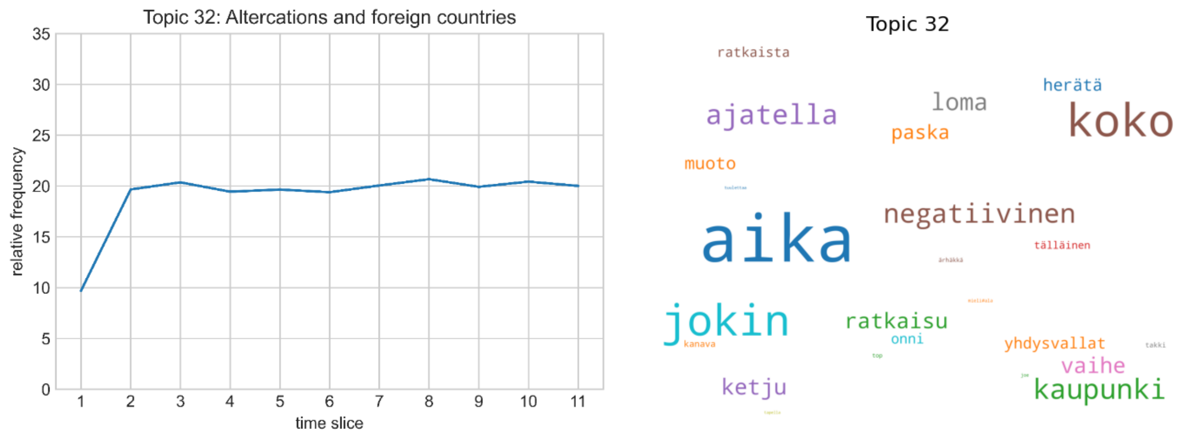


Figure 11. Temporal distribution and keywords for topic 32

With the exception of the very first time slice, the distribution of topic 32, concerning altercations and foreign countries, remains quite stable. This topic highlights the global nature of the crisis. Laypeople use events and developments in foreign countries as evidence of possible futures in Finland and may use these as supporting arguments for their own prescriptive analyses. In (4), a user responds to other users comparing COVID-19 to the flu. They use China's response to the virus as evidence that the disease is more serious than the common flu.

(4) @user1 @user2 Uskomatonta... flunssa? meneekö Kiinassa kaupungit kiinni ärhäkän flunssan takia? mitä varten valtion kanava näytti uutisista, jossa coronasta "selvinneet" tuulettavat onnesta soikeena. Ei kai tälläistä vakuutteluja tarvita, jos kyse olisi tavallisesta ärhäkästä flunssasta.

Other tweets in this topic are about the challenges and sacrifices that people face due to the pandemic. There is a general negative tone in the narratives told within this topic. Understandably, a crisis induces negative emotions in those who experience it. (5) is an example of a small story, told in two separate tweets. It tells a seemingly very mundane, personal story of waiting a long time on the telephone to receive a timeslot for a COVID-19 test. Still, the user has chosen to share the story, perhaps just to be part of the broader discussion, to show that they, too, are affected by the crisis.

(5) Uudella ekaluokkalaisella kurkku kipeä. Ohjeen mukaan pitää testata. Tässä 62minuuttia puhelimessa jo odottaneena ehtii kelailla. Tilanne on se ettei että ei

lapsi voi myöskään vaan jäädä kotiin varotoimena ilman testaamista.. Kuinkahan kauan tässä voi saada vielä odottaa? :) #koronafi

Update# 2: Läpi päästiin, kun aikaa oli kulunut 170minuuttia! Aika lauantaille! Over and out. Aika ruuhkaiselle vaikuttaa, sillä tuo lauantain aika oli juuri puhelun aikana vapautunut peruutusaika. #koronafi

4.5.4 Topic 22: Sweden and border controls

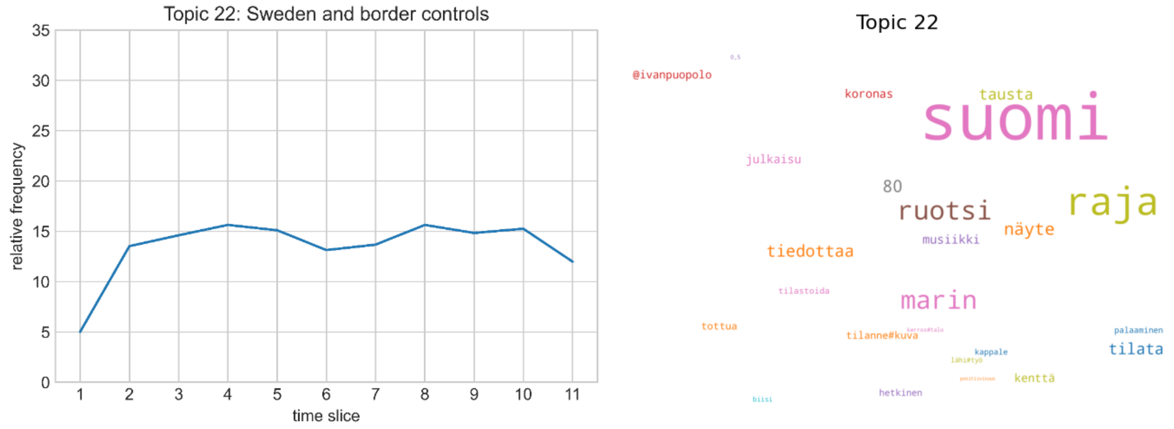


Figure 12. Temporal distribution and keywords for topic 22

Under normal circumstances, travelling between EU-countries can seem no different than travelling within one country, due to lax border formalities. Therefore, sudden restrictions on who is allowed to cross a border and for what reason imposed as COVID-19 began spreading in Europe came as a shock. Topic 22 contains discourses about the topic of border controls, especially between Finland and Sweden. Travelling between Finland and Sweden is commonplace, especially in places like Haaparanta and Tornio, two towns on different sides of the border, yet socially and economically closely linked. In (6), a user expresses sadness over the decision to close the border between these two towns.

(6) Suomen ja Ruotsin raja suljettiin Torniossa ja Haaparannalla. Vaikka raja suljetaan hyvästä syystä, se tuntuu surulliselta. Kävin aamulla kuvaamassa rajalla, kuvasarja blogissa: [link] #koronasuomi #tornio #haaparanta #rajakiinni #raja

Sweden gained some notoriety around the world for restraining from enforcing strict regulations and lockdowns. While in most of Europe restaurant and schools were closed, Swedish businesses remained mostly open. This attracted both praise and ire from the international audience. The two very different policies chosen by Finland and Sweden, and the effect they might have on each other, rises to prominence within this topic, as exemplified in (7). In this example, the user blames Sweden and Swedes for their own troubles and paints an exaggerated picture of Swedish refugees fleeing to Finland. They express a similar attitude of 'us versus them' as above, in (2).

(7) Jos Ruotsin itse aiheutettu koronakaaos jatkuu, meillä on ennen pitkää Lapin rajoilla tungokseen asti koronaa pakenevia ruotsalaisia. "Ruotsi ja Norja eivät halua Suomen sulkevan Lapin rajaa?", Jahas. Asiasta nyt kuitenkin päättää Suomi. #koronafi

4.5.5 Topic 0: Fear of death and sickness

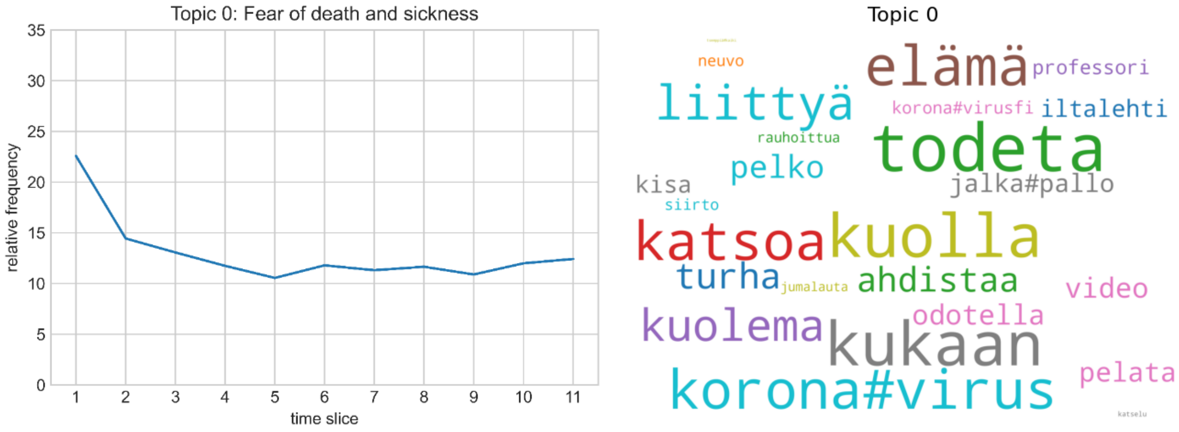


Figure 13. Temporal distribution and keywords for topic 0

Topic 0 covers discourse concerning a fear of getting infected by the virus and dying because of it. In time slice 1, few people had contracted the disease, but speculation about its spread and lethality was common on Twitter, with topic 0 among the most relevant topics in more than 20 % of tweets. Tweets written in time slice 1 reflect an anxiety of what is to come. With little hard data about the virus available, narratives comparing the virus to the common cold and narratives warning about the danger of the virus arose, as exemplified in (8) and (9).

(8) @user Koronavirus aiheuttaa hengitystieinfektion. Flunssaan kuolee maailmalla vuosittain 2-3 miljoonaa ihmistä. Montako ei-kiinalaista on kuollut Kiinan ulkopuolella koronavirukseen? Alle 5? Eikai tässä vielä syytä paniikkiin. Flunssa-aallot menevät yleensä ohi.

(9) Karseeta! Koronavirus ei ole pikku flunssa, numerot kuolleista yms on pahasti alakantissa. Katsokaa tää jos vielä ehditte. Monesti tämmöset videot ajetaan alas lyhyen ajan sisällä.

This topic is particularly interesting, since it deals with the disease itself rather than society's response to the virus, which is the focus of most other topics. The frequency of this topic is highest before any regulations were enacted and gradually decreases during the first months of the crisis. This suggests that when the crisis properly began, discourse shifted from the disease to reactions to the disease.

Death is considered a tabu subject in our culture and most avoid thinking about their own mortality. COVID-19 forced the topic onto people's minds and into public discourse. A topic usually reserved for private conversations was suddenly brought into daylight, both as personal stories such as (10) and as daily death statistics published by news media and shared by private individuals. Such practices seem grim in retrospect but are explained by the broader crisis narrative that allowed for and was partly created by the sharing of these statistics.

(10) kuolemaan ja kuinka se lähinnä ahdistaa ja herättää pelon tunteita. Havahduin siihe kuinka etäinen asia kuolema minulle on. Kyseisen ikimuistoisen kohtaamisen jälkee ja koronaepidemian myötä olen ajatellut enemmän kuolemaa. Ja havahtunut elämän rajallisuuteen ja ainutlaatuisuuteen.

4.5.6 Topic 83: Public spaces and services

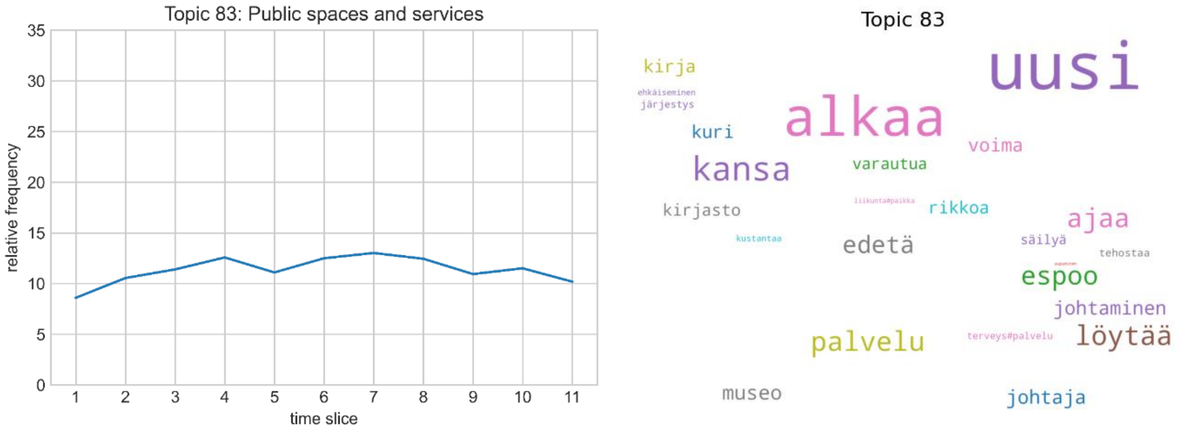


Figure 14. Temporal distribution and keywords for topic 83

At times during the pandemic, libraries, museums, and other public spaces were either closed or their use was severely limited. The Finnish are often characterised as book-loving people and there is some evidence showing they read more than the average European (Eurostat, 2018). Topic 83 seems to reflect that love for books and libraries. In many tweets, such as (11), libraries were seen as an important part of society. Online services were provided when physical locations were closed. Within this topic, there is present a worry about the long-term effects of closures on people's mental well-being.

> (11) Kirjastopalvelujen merkitys korostuu poikkeusoloissa. Vaikka kirjastotilat ovat nyt kiinni, moni palveluista verkossa on auki. Palveluita ei saisi nyt ajaa alas. @kirjastoseura #kirjasto #kirjastot #korona #koronavirus #jyty

4.5.7 Topic 64: Development and spread of pandemic

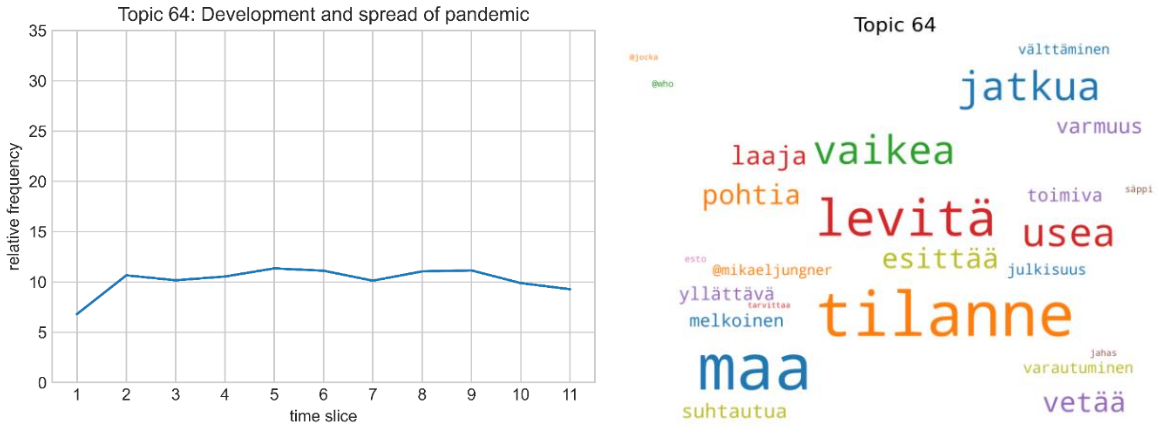


Figure 15. Temporal distribution and keywords for topic 64

Topic 64 covers discussion about how and where the virus spread, both within Finland and internationally. Alternative expertise, which was noted to be a major trend in Twitter discussion relating to COVID-19 in Väliverronen et al. (2020) is particularly present in this topic. Example (12) shows a user accusing THL of continually underestimating the seriousness of the COVID-19 disease based on a WHO report in the early stages of the pandemic. By questioning the local authority, the user creates a narrative that they cannot be trusted. The use of the hashtag '#hallitus' (*#government*) is read as a plea for the government to also disregard the local authority and to act decisively.

(12) THL:n #koronavirus-vähättely jatkuu: (01:58:30) @WHO toteaa korona'viruksen: 1. Aiheuttavan influenssaa vakavamman sairauden 2. Olevan influenssaa kuolettavampi #hallitus #helsinki #influenssa #flunssa

On the other hand, real experts also took part in the conversation. Lasse Lehtonen, a leading expert at Helsinki University Hospital (HUS) and going by the username lasleh on Twitter was notably active participant in public discourse both in traditional media and on Twitter. As noted in section 4.1, he was the third most active user in this dataset, contributing over 2,000 tweets. (13) is an example of him taking part in the crisis narrative by expressing concern about the volatile situation in the Helsinki metropolitan are in late 2020.

(13) Koronataistelun vaikeat viikot ovat nyt. Ratkaisevaa on pääkaupunkiseudun tilanne. Suuri väkimäärä tekee jäljittämisestä ja tukahduttamisesta vaikeaa. Vakava paikallinen epidemia voi käynnistyä äkkiä ja levitä sitten muualle maahan #covid19 #Uusimaa

4.5.8 Topic 4: Testing and symptoms

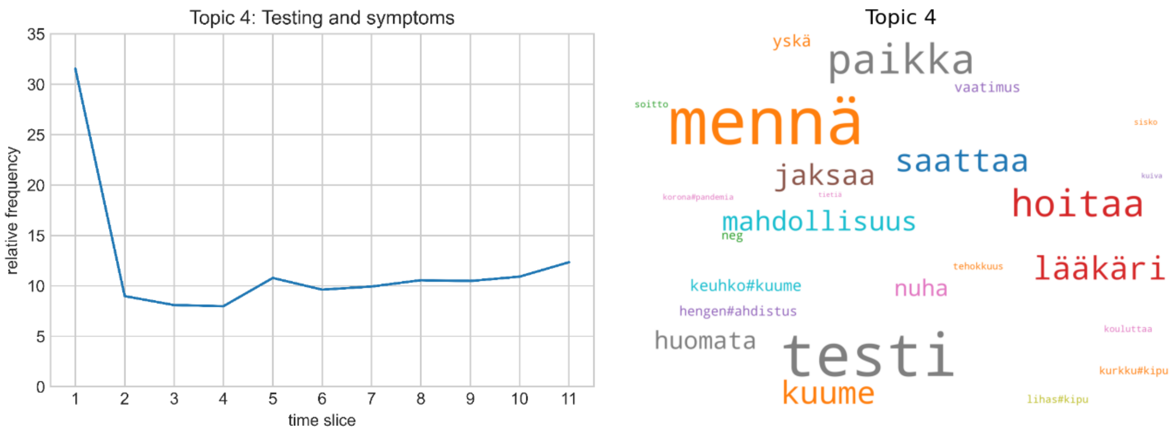


Figure 16. Temporal distribution and keywords for topic 4

Topic 4 has a sharp peak in frequency in the first time slice. As this time slice represents the primordial stage of crisis, it contains much discourse and speculation about what is to come. In the case of this topic, many tweets express anxiousness about how the pandemic will affect themselves. In (14), a user reports that they and their family are showing some symptoms. They ask how the situation is in Finland and wonder is the symptoms could be caused by COVID-19. As is typical to the primordial stage, there is a sense of unease and anticipation in this tweet: something is about to happen, but what and when exactly, is a mystery.

(14) Mites nyt Suomessa tämä #koronavirus etenee? Lähipiirissä paljon flunssa oireita. Kuumetta ja limaista yskää. Itsellä ollut nuhaa jo toista viikkoa. Ettei Että ei vaan ny olis kuitenkin samaa?

Discussion early on in the crisis was focused on identifying the symptoms of the disease. Many users on Twitter informed others of the most typical symptoms. Additionally, users also shared which symptoms they themselves were experiencing. The common flu and influenza were often used as a point of reference to compare symptoms to. Most tweets in this topic emphasize the difference between the symptoms between the flu and COVID-19, such as the example in (15). The general narrative is that COVID-19 is not the same as the flu and should be taken seriously.

(15) @user Jos on oireet: kuume, kuiva yskä ja hengenahdistusta, todennäköisesti koronaa. Influenssaan kuuluu nuha.

4.5.9 Topic 24: Work life

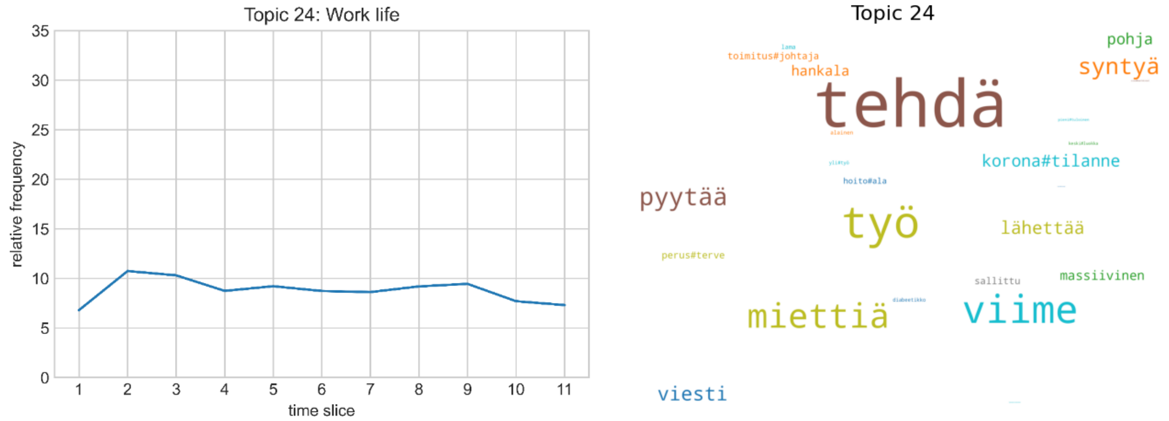


Figure 17. Temporal distribution and keywords for topic 24

The pandemic caused a massive change in the way many people worked. Working exclusively from home, a previously rather unusual phenomenon, suddenly became the norm for many people. This caused much discourse on Twitter. Both personal narratives and broader socioeconomic analysis were reported on the platform. Most tweets have paint a negative picture of working from home or criticize authorities for their handling of the pandemic. This topic peaks in frequency in time slice 2, at the start of the crisis. Another slight peak happens in time slice 9, when new infections were on the rise and the government was planning to enact restrictions on movement.

Although many people did work from home during the pandemic, this was not an option to people working in industries, where physical presence at work was required. This prompted some commentors to raise concern over the safety of the people, who could not stay home in isolation. In example (16), a user calls for these people to be vaccinated as soon as possible, due to the nature of their work.

> (16) Yhteiskunnan etulinjan ammattiryhmiä pitäisi rokottaa nopeasti. Päiväkotien ja koulujen henkilökunta, poliisit, sosiaalihuollon ammattilaiset. Huomennakin tätä työtä tehdään riskillä. Eikä tätä työtä tehdä etänä. #korona #työ

4.5.10 Topic 77: Government briefings

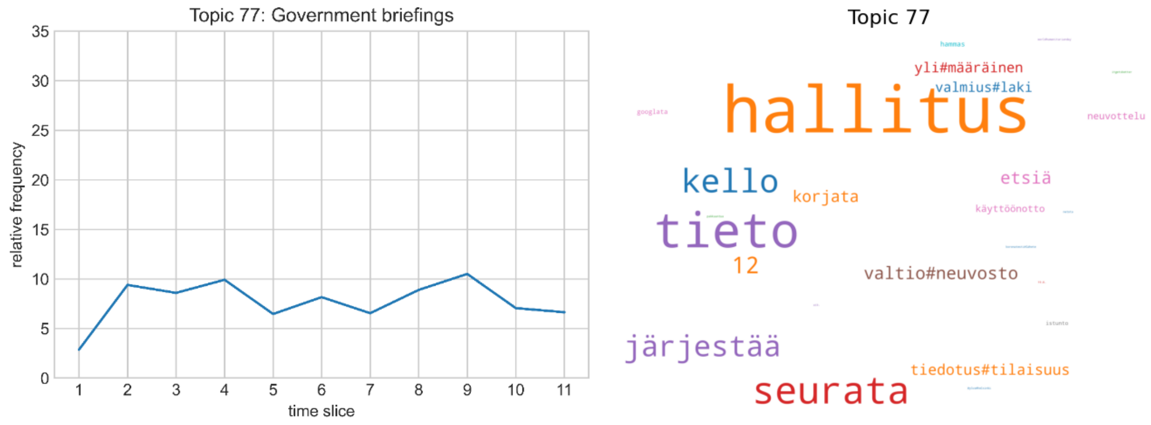


Figure 18. Temporal distribution and keywords for topic 77

Regular government briefings on updates regarding the development of the pandemic and changes in regulations were commonplace during the pandemic. This topic contains mainly tweets sent by accounts of official government bodies and traditional news media. This differentiates this topic from other topics discussed above. While other topics largely consist of individuals sharing their experiences and thought about the crisis, this topic consists mostly of crisis control and management. This topic shows that Twitter was used as one avenue of reaching people for authorities. Through Twitter, the public was informed about upcoming government briefings and the status of current restrictions. Other authorities used Twitter to inform users about possible places of exposure to the virus, such as example (17), sent by the account of the city of Kuopio.

> (17) - Puuilo klo 10-10.40 - K-market Pirtinportti klo 16-18 - R-kioski Kasarmikatu klo 16.15-16.50 - Satamaan K-market klo 20.30-20.35 - Ravintola Albatrossi klo

> 21-21.30 - Ravintola Sataman Helmi klo 21.30-23.00 Oireisena➡testiin: #kuopio #korona # koronavirus

There are some notable peaks in the trajectory of the topic over time. The first peak occurs in time slice 2, when the Emergency Powers act was enacted. Another peak occurs in time slice 4, which corresponds to a period of lessening restrictions and the withdrawal of the Emergency Powers act. Finally, in time slice 9, the topic peaks again. This period saw a sharp rise in new infections and discussion about limiting freedom of movement.

5 Discussion

In this section, I further discuss how the results of this study reflect on the crisis narrative regarding the COVID-19 pandemic. While in the results section the most frequent topics were discussed mostly in isolation, in this section, I draw connections between them, highlighting similarities and differences. To get a cohesive understanding of the crisis narrative, the topics must be seen as they are: simultaneous, interactive, and overlapping discourses.

5.1 Users and hashtags

There is no one account type that dominated the discussion. Among the most active users, there was a wide variety of user types. Even though the median user only sent two tweets, and a handful of users were particularly active, the bulk of all tweets were written by relatively inactive, private users. The crisis narrative that emerged was truly collaborative. Still, the more active a user is, the more chances they have on getting their stories seen by a large audience. Thus, active individuals with large audiences may have a stronger effect on the emerging meta-narrative. This is why it is so important to examine individual narratives and not only look at the data from afar, as numbers and statistics. Such an approach would obscure the peculiarities and intricacies of smaller stories.

What is striking, is that the most active users in this data do not include major news media or central political figures. Although authorities used Twitter to reach the public, central political figures seem to have kept mostly quiet on their personal accounts. Their effect on the crisis narrative on Twitter was mainly mediated through others. However, there are two politicians on the list of most active users. They are only minor figures in Finnish politics, unknown to most people. Their active participation in the discussion on Twitter can be seen as an attempt to gain exposure.

The most frequent hashtags largely correspond to the most frequent topics. This is to be expected since hashtags can be used as topic indicators. This also shows that the topics produced by the topic model accurately represent the data. Popular hashtags relate to politics (‘hallitus’, ‘talous’, ‘politiikka’), healthcare and restrictions (‘thl’, ‘karanteeni’, ‘etätyö’, ‘pysykotona’, ‘rokote’) and locations (‘suomi’, ‘helsinki’). The most used hashtags by far were #korona and #koronafi. Both hashtags are very general, and thus unlikely to add necessary information, considering the societal context in which they were authored. Their main function seems to be to connect tweets containing these hashtags to the broader discourse and thereby contributing to the meta-narrative of crisis.

5.2 Topics and crisis narrative

The results show that aspects of the crisis that directly affected people's lives gained more attention that aspects that seemed more distant. Hence, restrictions and regulations were discussed more than the disease itself. The transformed lives in which people found themselves in prompted them to share their personal narratives about the hardships, small annoyances, and absurd circumstances they faced. Most tweets and topics analysed in this study were, indeed, inherently negative. Feelings of sadness (7) and frustration (6) were common. On the other hand, some users shared narratives with a jovial tone (1). This contrast between serious and jovial attitudes is present throughout the data. I argue that they represent competing narratives about how the crisis is felt and characterised. Both show different survival strategies to cope with the crisis; others approach the crisis with the seriousness they believe it deserves, while others stay optimistic and try to find the silver lining in a difficult situation. The strategies might both be used by the same people in different situations and time periods.

The method chosen for this study did not reveal much temporal variation in the distribution of the topics. Most topics remained stable throughout the data collection period, often with the exception of time slice 1, which is unique due to its size and its situation before the acute stage of the crisis began. Outside of time slice 1, there were some peaks in certain topics that corresponded with changes in restrictions or infection rates. Hence, although the crisis narrative remained relatively stable in terms of topic distirbution, changes in the real world did affect how the crisis was conceived of in narrative. The nature of crisis is cyclical, and a prolonged crisis like this one is revealed to have its own inner cycles of acute unrest and relative calm. This reveals Fink's models of crisis, although clear and illustrative, is far too simple to capture the more intricate evolutions of a real crisis. In this study, I have not given the 'resolution' label to any time slice, arguing that the crisis did not end during the data collection period. However, this claim is challenged in some tweets. In fact, when a crisis is truly over is unclear, and different people are ready to let go of the crisis narrative sooner than others. These users began putting forth counternarratives that called for a return to normal, while others tried to continue the narrative of an ongoing and active crisis.

Major news events, such as the isolation of Uusimaa from the rest of Finland, and the Delta variant brought to Finland by football fans, were present in the data. This shows that traditional media narratives and social media narratives are connected. The direction of information flow remains unclear: Do social media narratives emerge from traditional media, vice versa, or are both influenced by the other. Surprisingly, although masks and vaccination were much discussed topics in the media, they did not emerge from the data as their own topics. This is made even more surprising by the fact that they were used as search terms for reducing the original dataset from 1.3 million tweets to roughly 380,000. I suspect that this is due to the original search terms that were used in the data collection phase. These search terms included many terms related to sickness and symptoms and did not include terms related to vaccines or masks. Since the data were collected in real time, as it was produced, the search terms had to be decided in the early stages of the pandemic, when it was unclear how the crisis would evolve and what discourses would emerge. This is an issue inherent to all corpus analysis. How a corpus is constructed always affects the results and conclusions that are drawn from it.

Out of the ten most frequent topics, only topics 0 (fear of death and sickness) and 4 (testing and symptoms) were directly related to the disease and its immediate effects. Both topics were most common in time slice 1 and then dropped in frequency after the implementation of restrictions and the Emergency Powers act. After this, other topics became more prominent. My analysis shows that for most people, the COVID-19 crisis was not a health crisis but a crisis of work, free time, and travel. The virus itself did not affect as many people’s daily lives as did the measures to mitigate its spread and, therefore, these measures became the crisis.

I do not argue that the restrictions and regulations on people’s freedom that were put in place during the pandemic were excessive, nor that they were insufficient. That is a matter that will surely be discussed in numerous other works in the years to come. Instead, I argue that even though the COVID-19 disease was the primus motor and underlying cause for the crisis, it was not the focus of the crisis narrative. In narrative, the crisis became about the daily lives of people and how they were forced to change their patterns of behaviour in response to the pandemic. This I believe to be the most important finding of this research, and something that needs to be considered in future societal crisis events. Authorities and governments have to listen closely to the voices of their people, because in their narratives, crisis might manifest very differently than expected.

6 Conclusion

In this thesis, I have examined a corpus of tweets collected between January 2020 and August 2021 which was purposefully constructed to represent discussion surrounding the COVID-19 pandemic that took place during the collection period. First, I gave a summary of the major developments in the pandemic in Finland to contextualise the circumstances under which the data analysed in this study came into existence. I examined the theoretical frameworks of crisis, narrative and crisis narrative and showed how they can be used to analyse social media discourse. I also introduced some of the central tenets of Digital Discourse Analysis and Corpus-Assisted Discourse Studies. The quantitative method used in this study is LDA topic modelling. It is a well-known and proven method of finding latent topics in large text corpora, which would be impossible to discover using traditional methods of discourse analysis. The data were divided into time slices and the latent topics discovered using topic modelling were then ranked by frequency. The ten most frequent topics were then analysed within the theoretical framework of crisis narratives.

The research questions presented in the Introduction were:

(1) Who were the most active participants in creating crisis narratives on Twitter?

(2) What were the most common hashtags and how do they relate to the crisis narrative?

(3) What topics and narratives within those topics arose in Finnish Twitter discussions during the COVID-19 pandemic?

(4) How did the distribution of topics and the crisis narratives change during the course of the pandemic?

The results of the study show that while most users were relatively inactive, there were a number of user types that were very active. No singular user type dominated the discourse. The most common hashtags were directly linked to the pandemic and were thereby connected to the broader discourse. The hashtags were also reflected in the most common topics, supporting the legitimacy of the topic model. The results show that the crisis narrative revolved mostly around the restrictions and regulations and not the disease itself. Topics 50 (‘restrictions and regulations’) and topic 26 (‘quarantine’) were particularly frequent throughout the crisis. I argue that this is because while everyone noticed the changes they had to make in their personal lives to accommodate the new rules, fewer people personally experienced the effects of the virus. While there was some temporal variation in the frequency of topics, these were mostly quite minor. Still, spikes often correlated with changes in the regulations or new infections. Alternative methods of tracking topics in time could have produced different results and shown greater variation in topic frequencies.

This study is another testament to the efficacy of topic modelling as a reliable and valuable tool for discourse analysts. Combining the quantitative method of topic modelling and the qualitative methods of more traditional discourse analysis enabled both a detailed analysis of individual narratives and a look at the bigger picture. There is still work to be done to improve topic modelling as a method for discourse analysis. Better and more robust model validation methods are direly needed. The coherence measure used to validate the model in this study is only one of many and it is unclear, which method most reliably predicts highly coherent topics. Each method uses different ways to calculate the coherence score, inherently carrying with them their own biases.

This study contributes to the understanding of crisis discourse in general and especially on social media. Whereas previous studies on the language of crisis focus primarily on issues of management, crisis control and corporate communication, this study examines crisis from a less studied perspective. This study brings into forefront the viewpoint of those living though and directly suffering from the effects of crisis. There is a great need to understand crisis from this perspective as mankind prepares to face future calamities. In order to understand how future crises unfold and how to best prepare for them, it is paramount to understand how they are constructed in narratives.

This study has shown that crisis is narrativized both as personal stories and as broader narratives. Some narratives examined in this study tell about feelings, personal struggles, and short moments in people’s lives, while others discuss national and international policies. Still, the topics that receive most attention are the ones that affect people personally. It is also evident that a crisis is multifaceted and ever evolving. It is cyclical both in the sense that one crisis leads to another but also internally, with moments of rest followed by great turmoil. General models can be useful to conceptualise crisis, but they can hardly be used to accurately describe the intricacies of real crises.

Future research could use the same data used in this thesis to examine crisis narratives on an individual level. Whereas this thesis examined crisis narratives and their development on a wider scale, future research could choose to focus on the most prolific users identified here and analyse how narratives told by them evolved during the pandemic. Additionally, a more intertextual study could look at the interplay between traditional news media and social media by analysing what and how news stories are shared and commented. Finally, other topic modelling methods, such as dynamic topic modelling or structural topic modelling could reveal more about the structure of the data.

References

Arun, R., Suresh, V., Veni Madhava, C. E., & Narashima Murty, M. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In M. J. Zaki, J. Xu Yu, B. Ravindran, & V. Pudi (Eds.), *Advances in Knowledge Discovery and Data Mining*. Springer.

Au, L., & Eyal, G. (2022). Whose Advice is Credible? Claiming Lay Expertise in a Covid-19 Online Community. *Qualitative Sociology*, *45*, 31–61. https://doi.org/10.1007/s11133-021-09492-1

Baron, N. S. (2008). *Always On : Language in an Online and Mobile World*. Oxford University Press.

Benson, P. (2015). YouTube as Text: Spoken Interaction Analysis and Digital Discourse. In R. H. Jones, A. Chik, & C. A. Hafner (Eds.), *Discourse and Digital Practices: Doing Discourse Analysis in the Digital Age*. Routledge.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic Topic Models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120. https://doi.org/10.1145/1143844.1143859

Blei, D. M., NG, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4/5), 993–1022.

Chang, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems*, 288–296.

Cheng, Y., & Cameron, G. (2017). The Status of Social-Mediated Crisis Communication (SMCC) Research. In L. Austin & Y. Jin (Eds.), *Social Media and Crisis Communication*. Routledge.

Crystal, D. (2011). *Internet Linguistics: A Student Guide*. Routledge.

Danet, B., & Herring, S. C. (Eds.). (2007). *The Multilingual Internet: Language, Culture and Communication*. Oxford University Press.

Dayter, D. (2015). Small stories and extended narratives on Twitter. *Discourse, Context and Media*, *10*, 19–26. https://doi.org/10.1016/J.DCM.2015.05.003

De Fina, A., & Georgakopoulou, A. (2011). *Analyzing Narrative: Discourse and Sociolinguistic Percpectives*. Cambridge University Press.

De Rycker, A., & Mohd Don, Z. (2013). Discourse in crisis, crisis in discourse. In A. De Rycker & Z. Mohd Don (Eds.), *Discourse and Crisis: Critical Perspectives* (pp. 3–65). John Benjamins Publishing Company.

Eurostat. (2018). *World Book Day*. https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20180423-1

Fink, S. (1986). *Crisis Management: Planning for the Inevitable*. American Management Association.

Finlex. (2020). *Valtioneuvoston asetus valmiuslain 86, 88, 93–95 ja 109 §:ssä säädettyjen toimivaltuuksien käyttöönotosta*. https://finlex.fi/fi/laki/alkup/2020/20200125

Finnish Government. (2020). *Restrictions on movement to and from Uusimaa enter into force on 28 March 2020*. https://valtioneuvosto.fi/en/-//10616/liikkumisrajoitukset-uudellemaalle-voimaan-28-maaliskuuta-2020-klo-00-00

Finnish Government. (2021a). *Finland declares a state of emergency*. https://valtioneuvosto.fi/en/-//10616/finland-declares-a-state-of-emergency

Finnish Government. (2021b). *Government publishes plan and timetable for lifting COVID-19 restrictions, consultation round begins*. https://valtioneuvosto.fi/en/-//10616/government-publishes-plan-and-timetable-for-lifting-covid-19-restrictions-consultation-round-begins

Finnish Institute for Health and Welfare. (2020). *Matkailijalla todettu koronavirustartunta Lapin keskussairaalassa*. https://thl.fi/fi/-/matkailijalla-todettu-koronavirustartunta-lapin-keskussairaalassa

Finnish Institute for Health and Welfare. (2021). *Tartuntatautirekisterin COVID-19-tapaukset [COVID-19 cases in the Infectious Diseases Register]*. https://sampo.thl.fi/pivot/prod/fi/epirapo/covid19case/fact_epirapo_covid19case

Finnish National Board on Research Intergrity TENK. (2019). Ihmiseen kohdistuvan tutkimuksen eettiset periaatteet ja ihmistieteiden eettinen ennakkoarviointi Suomessa: Tutkimuseettisen neuvottelukunnan ohje 2019. In *Tutkimuseettisen neuvottelukunnan julkaisuja*. https://tenk.fi/sites/default/files/2021-01/Ihmistieteiden_eettisen_ennakkoarvioinnin_ohje_2020.pdf

Garimella, K., Weber, I., & De Choudhury, M. (2016). Quote RTs on Twitter: Usage of the new feature for political discourse. *WebSci 2016 - Proceedings of the 2016 ACM Web Science Conference*, 200–204. https://doi.org/10.1145/2908131.2908170

Georgakopoulou, A. (2006). Thinking Big with Small Stories in Narrative and Identity Analysis. *Narrative Inquiry*, *16*(1), 122–130.

Georgakopoulou, A. (2017). Small Stories Research: A Narrative Paradigm for the Analysis of Social Media. In L. Sloan & A. Quan-Haase (Eds.), *The SAGE Handbook of Social Media Research Methods* (pp. 266–281). Sage.

Georgalou, M. (2015). Small Stories of the Greek Crisis on Facebook. *Social Media and Society*, *1*(2), 1–15. https://doi.org/10.1177/2056305115605859

Goodwin, C. (1984). Notes on Story Structure and the Organization of Participation. In J. M. Atkinson & J. Heritage (Eds.), *Structure of Social Action: Studies in Conversation Analysis* (pp. 225–246). Cambridge University Press.

Haines, A., & Ebi, K. (2019). The Imperative for Climate Action to Protect Health. *The New England Journal of Medicine*, *380*(3), 263–273. https://doi.org/10.1056/NEJMra1807873

Hay, C. (1996). Narrating Crisis: The Discursive Construction of the "Winter of Discontent." *Sociology*, *30*(2), 253–277. https://doi.org/10.1177/0038038596030002004

Heath, R. L. (1994). *Management of Corporate Communication: From Interpersonal Contacts To External Affairs*. Routledge.

Heath, R. L. (2013). Telling a Story: A Narrative Approach to Communication During Crisis. In D. P. Millar & R. L. Heath (Eds.), *Responding to Crisis: A Rhetorial Approach to Crisis*

Communication. Routledge.

Herring, S. C. (1996). *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. John Benjamins Publishing Company.

Herring, S. C. (2007). A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet*, *1*, 1–37.

Huang, M. (2020). Introduction: Constructing and Communicating Crisis Discourse from Cognitive, Discursive and Sociocultural Perspectives. In M. Huang & L.-L. Holmgreen (Eds.), *The Language of Crisis: Metaphors, Frames and Discourses*. John Benjamins Publishing Company.

Huang, R. (2019). Network fields, cultural identities and labor rights communities: Big data analytics with topic model and community detection. *Chinese Journal of Sociology*, *5*(1), 3–28. https://doi.org/10.1177/2057150X18820500

Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, *22*(5), 469–485. https://doi.org/10.1080/13645579.2019.1576317

Janmohamed, K., Soale, A., Forastiere, L., Tang, W., Sha, Y., Demant, J., Airoldi, E., & Kumar, N. (2020). Intersection of the Web-Based Vaping Narrative With COVID-19: Topic Modeling Study. *Journal of Medical Internet Research*, *22*(10). https://doi.org/10.2196/21743

Johansson, M., Kyröläinen, A.-J., Ginter, F., Lehti, L., Krizsán, A., & Laippala, V. (2018). Opening up #jesuisCharlie anatomy of a Twitter discussion with mixed methods. *Journal of Pragmatics*, *129*, 90–101. https://doi.org/10.1016/j.pragma.2018.03.007

Jones, R. H., Chik, A., & Hafner, C. A. (2015). Introduction: Discourse Analysis and Digital Practices. In R. H. Jones, A. Chik, & C. A. Hafner (Eds.), *Discourse and Digital Practices: Doing Discourse Analysis in the Digital Age* (pp. 1–17). Routledge.

Kanerva, J., Ginter, F., Miekka, N., Leino, A., & Salakoski, T. (2018). Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 133–142.

King, B. W. (2015). Investigating digital sex talk practices: A reflection on corpus-assisted discourse analysis. In R. H. Jones, A. Chik, & C. A. Hafner (Eds.), *Discourse and Digital Practices: Doing Discourse Analysis in the Digital Age* (pp. 130–143). Routledge. https://doi.org/10.4324/9781315726465-9

Knight, D. (2015). e-Language: Communication in the Digital Age. In A. McEnery & P. Baker (Eds.), *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Palgrave Macmillan.

Labov, W. (1972). *Language in the Inner City: Studies in the Black English Vernacular*. University of Pennsylvania Press.

Lehti, L., Luondonpää-Manni, M., Jantunen, J. H., Kyröläinen, A.-J., Vesanto, A., & Laippala, V. (2020). Commenting on poverty online : A corpus-assisted discourse study of the Suomi24 forum. *SKY Journal of Linguistics*, *33*, 7–47.

Li, X., & Lei, L. (2021). A bibliometric analysis of topic modelling studies (2000-2017). *Article Journal of Information Science*, *47*(2), 161–175. https://doi.org/10.1177/0165551519877049

McEnery, A., & Baker, P. (2015). Introduction. In A. McEnery & P. Baker (Eds.), *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Palgrave Macmillan. https://web.p.ebscohost.com/ehost/ebookviewer/ebook/bmxlYmtfXzEwNTQxNzBfX0FO0?sid=913f3eaa-c13b-423c-b4c3-6a39db328e77@redis&vid=0&format=EB&rid=1

Mutanga, M. B., & Abayomi, A. (2020). Tweeting on COVID-19 Pandemic in South Africa : LDA-Based Topic Modelling Approach. *African Journal of Science, Technology, Innovation and Development*, *14*(1), 161–172. https://doi.org/10.1080/20421338.2020.1817262

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2017). Topic modelling for qualitative studies. *Journal of Information Science*, *43*(1), 88–102. https://doi.org/10.1177/0165551515617393

Ochs, E., & Capps, L. (2001). *Living Narrative: Creating Lives in Everyday Storytelling*. Harvard University Press.

Official Statistic of Finland. (2020). *Seuratut yhteisöpalvelut 2020, %-osuus väestöstä [Social Media Use 2020, percentage of population]*. http://www.stat.fi/til/sutivi/2020/sutivi_2020_2020-11-10_tau_025_fi.html

Ojala, M., Pantti, M., & Laaksonen, S. (2019). Networked publics as agents of accountability : Online interactions between citizens , the media and immigration officials during the European refugee crisis. *New Media & Society*, *21*(2), 279–297. https://doi.org/10.1177/1461444818794592

Page, R. (2015). The Narrative Dimensions of Social Media Storytelling: Options for Linearity and Tellership. In A. De Fina & A. Georgakopoulou (Eds.), *The Handbook of Narrative Analysis* (pp. 329–347). John Wiley & Sons.

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and Meanings in Discourse: Theory and Practice in Corpus-assisted Discourse Studies (CADS)*. John Benjamins Publishing Company.

Pattamawan, J., & Todd, R. W. (2013). Red or Yellow, Peace or War: Agonism and Antagonism in Online Discussion During the 2010 Political Unrest in Thailand. In A. De Rycker & Z. Mohd Don (Eds.), *Discourse and Crisis: Critical Perspectives*. John Benjamins Publishing Company.

Phoenix, C., Smith, B., & Sparkes, A. C. (2010). Narrative Analysis in Aging Studies: A Typology for Consideration. *Journal of Aging Studies*, *24*(1), 1–11. https://doi.org/10.1016/j.jaging.2008.06.003

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Roberts, M. E., Stewart, B. M., Tingley, J., & Airoldi, E. M. (2013). The Structural Topic Model and Applied Social Science. *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*, 2–5.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *WSDM ’15: Proceedings of the Eighth ACM International Conference on Web Search and Data*

Mining, 399–408.

Schneider, K. T., & Carpenter, N. J. (2020). Sharing #MeToo on Twitter: Incidents, Coping Responses, and Social Reactions. *Equality, Diversity and Inclusion*, *39*(1), 87–100. https://doi.org/10.1108/EDI-09-2018-0161

Scollon, R. (2001). *Mediated Discourse: The Nexus of Practice*. Routledge.

Shahi, G. K., Dirkson, A., & Majchrzak, T. A. (2021). An Exploratory Study of COVID-19 Misinformation on Twitter. *Online Social Networks and Media*, *22*, 1–16. https://doi.org/10.1016/j.osnem.2020.100104

Siapera, E. (2014). Tweeting #Palestine: Twitter and the mediation of Palestine. *International Journal of Cultural Studies*, *17*(6), 539–555. https://doi.org/10.1177/1367877913503865

Stott, P. (2016). How climate change affects extreme weather events. *Science*, *352*(6293), 1517–1518. https://doi.org/10.1126/science.aaf7271

Thurlow, C., & Mroczek, K. (2011). Introduction: Fresh Perspectives on New Media Sociolinguistics. In C. Thurlow & K. Mroczek (Eds.), *Digital Discourse: Language in the New Media*. Oxford University Press.

Twitter. (2020). *Investor Fact Sheet*. https://s22.q4cdn.com/826641620/files/doc_financials/2018/q3/TWTR-Q3_18_InvestorFactSheet.pdf

Usher, K., Durkin, J., & Bhullar, N. (2020). The COVID-19 pandemic and mental health impacts. *International Journal of Mental Health Nursing*, *29*, 315–318. https://doi.org/10.1111/inm.12726

Vainikka, E., & Huhtamäki, J. (2015). Tviittien politiikkaa – poliittisen viestinnän sisäpiirit Twitterissä. *Media & Viestintä*, *38*(3), 165–183. https://journal.fi/mediaviestinta/article/view/62081

Väliverronen, E., Laaksonen, S.-M., Jauho, M., & Jallinoja, P. (2020). Liberalists and Data-Solutionists: Redefining Expertise in Twitter Debates on Coronavirus in Finland. *Journal of Science Communication*, *19*(5), 1–21. https://doi.org/10.22323/2.19050210

van Dijck, J., & Alinead, D. (2020). Social Media and Trust in Scientific Expertise: Debating the Covid-19 Pandemic in The Netherlands. *Social Media and Society*, *6*(4). https://doi.org/10.1177/2056305120981057

Walton, S., & Jaffe, A. (2011). "Stuff White People Like": Stance, Class, Race, and Internet Commentary. In Crispin Thurlow & K. Mroczek (Eds.), *Digital Discourse: Language in the New Media*. Oxford University Press.

WHO. (2020a). *Novel Coronavirus (2019-nCoV) Situation Report - 13* (Issue February).

WHO. (2020b). *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

WHO. (2021). *33rd WHO Regulatory Update on COVID-19* (Issue September 21).

Wicke, P., & Bolognesi, M. M. (2020). Framing COVID-19: How We Conceptualize and Discuss the Pandemic on Twitter. *PLoS ONE*, *15*(9), 1–24. https://doi.org/10.1371/journal.pone.0240010

Wikström, P. (2014). #srynotfunny: Communicative functions of hashtags on twitter. *SKY Journal of Linguistics*, *27*, 127–152.

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of Medical Internet Research*, *22*(11), 1–14. https://doi.org/10.2196/20550

Yang, T.-I., Torget, A. J., & Mihalcea, R. (2011). Topic Modeling on Historical Newspapers. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities 2011 Association for Computational Linguistics*, *June*, 96–104.

Zappavigna, M. (2013). *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. Bloomsbury Publishing.

Zappavigna, M. (2015). Searchable talk: the linguistic functions of hashtags. *Social Semiotics*, *25*(3), 274–291. https://doi.org/10.1080/10350330.2014.996948

Zappavigna, M. (2017). Twitter. In C. R. Hoffman & W. Bublitz (Eds.), *Pragmatics of Social Media*. De Gruyter Mouton.

Appendices

Appendix 1 Original search term list

#koronafi, #koronakriisi, #covid19fi, #pysykotona, #korona, #koronavirus, #koronasuomi, #koronaFI, #koronavirusFi, #COVID19, #poikkeustila, #karanteeni, #StayHome, #poikkeusolot, #koronavinkit, #koronakevät, #covid19finland, apua, apteekki, ahdistaa, ahdistus, kivussa, kipu, parantua, parantuminen, parantuneeni, sairaana, sairas, kotona, oireet, oireita, keuhkot, keuhkoissa, karanteeni, karanteenissa, hiki, lämpöä, kuume, yskä, hengenahdistus, lihaskipu, väsymys, nuha, päänsärky, vatsakipu, ripuli, särky, pääkipu, lihassärky, keuhkokuume, kipeä, kuumeilu, kurkkukipu, ruokahaluttomuus, vatsakivut, tukkoisuus, flunssa, pahoinvointi, yskää, vatsavaivat, kuumetta, yskiminen, hengitysvaikeuksia, uupumus, väsynyt, uupunut, aivastelu, korvasärky, köhä, niistäminen, nuhakuume, mahakivut, rasitus, tulehdus, koronteeni, koronapeli, etäsuomenyliopisto, koronakoulu, koronalle, korontreeni, korontreenit, korontai, covidiootti, karantini, kotonavitus, epäopetus, epäopiskelu, epäoppiminen, karantiimi, koronalinko, CoronavirusFrance, COVID19france, Quarantine, lockdown, restecheztoi, #coronase, #coronaswe, #coronaviruset, #coronavirussverige, #coronasverige, #coronasweden, #coronavirusSE, #coronakris, #coronakrisen, #coronavirussweden, #covid19sverige, #covid19sweden, #covid19se, #coronavirusfinland, #stannahemma, #coronafinland, #undantagstillstånd, #karantän, hjälp, apotek, ont, sjuk, smärta, bättra, bättring, friskna, hemma, symtom, lungor, karantän, feber, febrig, stegring, temperatur, hosta, andning, andas, muskel, muskler, trött, förkyl, snuva, influensa, flunssa, krasslig, verk, diareé, inflammation, aptit, ingen smak, inga smaker, tappat smak, tappat all smak, utan smak, matlust, ingen lukt, inga lukter, tappat lukt, utan lukt, täppt, täppa, mår illa, må illa, illamående, magbesvär, lös mage, utmattad, utmattning

Appendix 2 Suomenkielinen lyhennelmä

Johdanto

COVID-19 pandemia sai alkunsa vuoden 2019 lopussa Kiinan Wuhanissa, josta se levisi Eurooppaan ja koko maailmaan vuoden 2020 alussa. Ihmisten välisen fyysisen kanssakäymisen ollessa rajoitettua ja suositusten vastaista, internetin sosiaaliset mediat, kuten Twitter, mahdollistivat kriisistä keskustelun. Sosiaalisessa mediassa syntyi kriisinarratiivi, johon omalta osaltaan kuuluivat myös valeuutiset, auktoriteettien kyseenalaistaminen ja vaihtoehtoisen asiantuntijuuden kehittyminen (Au & Eyal, 2022; Väliverronen et al., 2020; van Dijck & Alinead, 2020). Tässä pro gradu -tutkielmassa tutkin Twitterissä jaettuja kriisinarratiiveja. Aineistonani käytän 375 322 Twitter-viestin eli tviitin korpusta, joka kerättiin Twitteristä vuoden 2020 ja vuoden 2021 ensimmäisen puoliskon aikana. Näin laajan tekstidatan analysointiin käytän topiikkimallinnusta. Topiikkimallinnus on laskennallinen metodi, joka laskee sanojen yhteisesiintyvyyden perusteella datasta ennaltamäärätyn määrän topiikkeja eli puheenaiheita. Suosituimpiin topiikkeihin pohjaten analysoin sekä yksittäisissä tviitteissä esiintyviä narratiiveja, että niiden muodostamaa laajempaa metanarratiivia. Tämän lisäksi tutkin, keiden käyttäjien tviittejä aineistossa esiintyy eniten, sekä suosituimpia *hashtageja* eli aihetunnisteita.

Tausta ja teoria

COVID-19 pandemian synnyttämää kriisinarratiivia voi ymmärtää vain, jos tuntee yhteiskunnallisen kontekstin, jossa narratiivi syntyi. Suomessa pandemia alkoi 29. tammikuuta 2020, kun ensimmäinen tartunta Suomessa varmistui. Tautitilanne pysyi rauhallisena, kunnes maaliskuussa tartuntatapaukset lähtivät voimakkaaseen nousuun ja hallitus julisti maahan poikkeusolot. Koulut suljettiin, ravintoloiden aukioloa rajoitettiin ja jopa Uudenmaan raja suljettiin väliaikaisesti viruksen leviämisen estämiseksi. Kesällä uusia tartuntatapauksia todettiin vähemmän, minkä johdosta monista rajoitteista luovuttiin. Syksyllä uusia tapauksia raportoitiin taas enemmän, ja uusia rajoitteita säädettiin. Tartuntatapaukset jatkoivat nousuaan ja maaliskuussa 2021 Suomi oli jälleen poikkeusoloissa. Loppukeväästä uusia tartuntoja raportoitiin taas vähemmän ja puhe exit-strategiasta alkoi.

Kriisinarratiivit

Kriisejä on tutkittu paljon, mutta suuri osa tutkimuksesta on tehty kriisinhallinnan ja kriisiviestinnän näkökulmasta (M. Huang, 2020). Kriisillä on monia määritelmiä ja piirteitä, joihin kuuluvat muun muassa negatiivisuus, muutos normaaliin ja äkillisyys (De Rycker & Mohd Don, 2013). COVID-19 pandemiaa voi siis hyvinkin pitää kriisinä, sillä se täyttää kaikki nämä määreet. Myös kriisien rakennetta on tutkittu. Fink (1986) on laatinut neliportaisen mallin, jossa kriisi kehittyy esivaiheesta akuuttiin kriisin, jota seuraa pitkittynyt kriisi ja lopulta kriisin päätös. Kriisien märitteleminen on vaikeaa, koska kriiseihin liittyy olennaisesti ennalta-arvaamattomuus ja jatkuva muutos. Jokainen kriisi on ainutlaatuinen tapahtuma, jolla on omat ominaispiirteensä. Tässä tutkimuksessa tarkastelen kuitenkin kriisejä erityisesti sosiaalisena ja kielellisenä ilmiönä. Tästä näkökulmasta katsottuna kriisit eivät ole ainoastaan aineellisen maailman tapahtumia, vaan myös ihmistenvälisestä diskurssista nousevia ilmiöitä. Kriisit syntyvät, kun media tai ihmisjoukko puhuvat kriisin todeksi luomalla kriisinarratiivin.

Moderni narratiivintutkimus sai alkunsa Labovin haastattelututkimuksissa (Labov, 1972). Labov havainnoi haastatteluissa kerrotuista tarinoista yhtenäisiä piirteitä, joiden pohjalta hän laati mallin tyypilliselle tarinalle. Labovin tutkimuksella oli suuri vaikutus yhteiskuntatieteisiin. Myöhemmin Labovin mallia on kuitenkin kritisoitu siitä, että se jättää huomiotta pienet tarinat (engl. *small stories*), jotka eivät noudata tyypillisen tarinan rakennetta tai kertovat meneillään olevista tapahtumista tai mahdollisista tulevaisuuden tapahtumista (Georgakopoulou, 2006). Labovilaista lähestymistapaa seuranneet keskusteluanalyyttiset näkökulmat tarkastelevat narratiivia keskustelijoiden välisen sanallisen ja sanattoman vuorovaikutuksen tuloksena. Toisaalta narratiivit voidaan nähdä myös tapana ymmärtää ja jäsennellä maailmaa. Viime vuosina suosiota on saavuttanut pienten tarinoiden tutkimus Georgakopoulou, 2006, 2017). Pienten tarinoiden tutkimus tarkastelee niitä narratiiveja, jotka syntyvät arkisessa keskustelussa. Pienet tarinat ovat usein lyhyitä, keskeneräisiä ja suunnittelemattomia (Georgakopoulou, 2006).

Kriisinarratiivit syntyvät Hayn (1996) mukaan, kun suuresta määrästä yksittäisiä tapahtumia ja tilastoja poimitaan uutiskynnyksen ylittävät tapaukset, jotka välitetään eteenpäin mediassa narratiivin muodossa. Näiden yksittäisten kriisinarratiivien muodostama kokonaisuus synnyttää laajemman metanarratiivin käynnissä olevasta kriisistä. Sosiaalisen median aikakautena kriisinarratiiveja kerrotaan ja jaetaan perinteisen median lisäksi myös esimerkiksi Twitterissä tai Facebookissa, kuten Georgalou (2015) ja Siapera (2014) ovat osoittaneet.

Digitaalinen diskurssianalyysi ja korpusavusteinen diskurssintutkimus

Tämä tutkimus nojaa vahvasti digitaalisen diskurssianalyysin ja korpusavusteisen diskurssintutkimuksen aloihin. Digitaalinen diskurssianalyysi, joka tunnetaan myös nimellä tietokonevälitteinen diskurssianalyysi, tutkii diskurssia digitaalisessa ympäristössä, kuten internetin keskustelupalstoilla ja sähköposteissa. Tekniikan nopean kehittymisen takia tutkimusala on jatkuvassa muutoksessa, minkä takia tutkimusalalla hyödynnetään monia teoreettisia viitekehyksiä ja metodeja (Jones et al., 2015). Yksi tutkimusalan peruspilareista on kuitenkin sen ymmärtäminen, että viestimiseen käytetty teknologia vaikuttaa viestin sisältöön. Digitaalisen diskurssianalyysin puutteena on monikielellinen tutkimus. Tällä hetkellä tutkitaan erityisesti englannin kielen piirteitä digitaalisessa ympäristössä muiden kielten kustannuksella (Thurlow & Mroczek, 2011).

Korpusavusteinen diskurssintutkimus yhdistää korpustutkimuksessa käytettyjä määrällisiä metodeja sekä perinteisen diskurssianalyysin laadullisia metodeja. Digitaalisen diskurssin tutkimiseen korpuksia on käytetty esimerkiksi kielen muodollisuuden vertailuun (Knight, 2015), digitaalisen seksipuheen tutkimiseen (King, 2015) sekä köyhyysdiskurssien tutkimiseen (Lehti et al., 2020).

Topiikkimallinnus

Topiikkimallinnus on laskennallinen metodi, jonka avulla on mahdollista tutkia laajoja tekstikokoelmia. Topiikkimallinnuksen avulla tekstikokoelmasta voi löytää tekstejä yhdistäviä topiikkeja eli puheenaiheita. Topikkimallinnusta on mahdollista tehdä monien eri mallien avulla, mutta suosituin ja tässäkin tutkimuksessa käytetty on nimeltään Latent Dirichlet Allocation (LDA).

LDA perustuu sille olettamalle, että jokainen sana tekstissä kuuluu johonkin topiikkiin. LDA laskee, kuinka usein samat sanat esiintyvät yhdessä tekstissä eli dokumentissa ja määrittää joka sanalle todennäköisyyden sille, että se kuuluu tiettyyn topiikkiin. Sanat, jotka esiintyvät usein yhdessä, kuuluvat todennäköisemmin samaan topiikkiin kuin sanat, jotka esiintyvät yhdessä vain harvoin. Mallinnuksen tuloksena syntyy tutkijan päättämä määrä topiikkeja, joita edustaa avainsanalista. Avainsanalista perusteella tutkija voi päätellä topiikin aiheen.

Topiikkimallinnusta on hyödynnetty niin sosiologiassa ((R. Huang, 2019), historiassa (Yang et al.,2011) kuin lingvistiikassakin (Jacobs & Tschötschel, 2019). Sitä on myös käytetty koronanpandemiaan liittyvien diskurssien tutkimiseen Twitterissä. Mutanga & Abayomi (2020) havaitsivat topiikkimallin avulla, että Etelä-Afrikassa Twitter ja perinteinen uutismedia käsittelivät samankaltaisia aiheita pandemian alussa. Wicke & Bolognesi puolestaan tutkivat pandemiasta käytettyjä metaforia. Heidän tutkimuksensa mukaan pandemiaa verrattiin eniten sotaan.

Twitterin kielellisiä piirteitä

Niin kuin kielellä kaikissa ympäristöissä, myös Twitterissä käytetyllä kielellä on omat ominaispiirteensä. Selkein Twitterin kieleen vaikuttava tekijä on 280 merkin raja, jota yksittäinen tviitti ei voi ylittää. Tämän lisäksi Zappavigna (2013) luettelee kolme tviittien kielen erityispiirrettä: @-merkin käyttö muihin käyttäjiin viitatessa, #-merkin käyttö aihetunnisteena, sekä uudelleentviittaaminen eli toisen käyttäjän viestin jakaminen uudestaan. Kommunikoinnin Twitterissä voi tulkita olevan eriaikaista, sillä toisin kuin kasvokkain tapahtuvassa kommunikaatiossa, keskustelijoiden ei tarvitse olla läsnä samanaikaisesti. Lisäksi viestintä Twitterissä on yhdensuuntaista, sillä viestin lukija voi reagoida vasta, kun viesti on täysin kirjoitettu ja lähetetty.

Aineisto ja menetelmät

Tämän tutkimuksen aineisto kerättiin Twitteristä hakusanalistan avulla. Aineiston keruu aloitettiin 1. huhtikuuta 2020 ja lopetettiin 6. elokuuta 2021. Tämän jälkeen kerättiin vielä jälkikäteen tviittejä, jotka oli lähetetty ennen aineiston keruun aloittamista. Aineiston keruussa käytetty hakusanalista on tutkimuksen liitteenä (Appendix 1). Kaikkiaan aineistoon kerättiin 1 339 416 tviittiä. Aineiston alkuperäinen käyttötarkoitus oli pandemian tosiaikainen seuraaminen, minkä takia hakusanalistassa oli paljon oireisiin liittyvää sanastoa. Tästä syystä aineistossa oli paljon tviittejä, jotka eivät olleet tämän tutkimuksen kannalta olennaisia ja ainestoa piti siivota. Ensin aineistosta poistettiin vuonna 2019 lähetetyt tviitit, uudelleentviittaukset sekä duplikaatit. Tämän jälkeen laadittiin uusi hakusanalista, jonka sanat liittyivät läheisesti koronapandemiaan ja sen ehkäisyyn. Lopullisessa aineistossa on 375 322 tviittiä.

Yhteensä aineistossa on 4 477 614 sanaa, ja yksi tviitti on keskimäärin 11,93 sanan mittainen. Kaikkein eniten tviittejä aineistossa on maalis- ja huhtikuulta 2020, 70 049 ja 71 970. Selvästi vähiten tviittejä on tammi- ja helmikuulta 2020, 81 ja 198. Tviittien epätasaisen jakauman vuoksi aineiston kuukausittainen vertailu on hankalaa. Tästä syystä aineisto jaetiin kahden kuukauden mittaisiin aikajaksoihin jakauman tasoittamiseksi. Ainoastaan vuoden 2020 maalis- ja huhtikuun tviitit analysoitiin yhden kuukauden aikajaksoissa.

Ennen topiikkimallinnusta tviitit lemmatisoitiin eli muutettiin perusmuotoonsa ja isot kirjaimet muutettiin pieniksi. Hyperlinkit, välimerkit ja yhden kirjaimen sanat poistettiin. Sanojen sanaluokat tunnistettiin automaattisesti ja konjunktiot, adverbit ja adpositiot poistettiin. Lopuksi poistettiin vielä sanat, jotka esiintyivät vain yhdessä dokumentissa tai vain yhden kerran koko aineistossa.

Topiikkimallinnus toteutettiin Gensim (Rehurek & Sojka, 2010) python-kirjaston avulla. Sopivan topiikkimäärän löytämiseen käytettiin C_V-koheressimittaa (Röder et al., 2015), jonka avulla mallin tulosten laatua pystyy arvioimaan numeerisesti. Koherenssimitan perusteella sopivaksi topiikkimääräksi valikoitui 110. Jokaista topiikkia edusti 25 avainsanaa. Topiikin avainsanat ovat sanoja, joilla on suurin todennäköisyys kuulua kyseiseen topiikkiin. Jokaiselle tviitille laskettiin viisi yleisintä topiikkia, minkä perusteella laskettiin yleisimmät topiikit koko aineistossa. Tämän jälkeen 21 yleisintä topiikkia nimettiin avainsanojen perusteella. Kaikista yleisin topiikki käsitteli yleistä korona- ja rajoitekeskustelua ja oli niin yleinen, että sen katsottiin kuvaavan aineistoa kokonaisuutena. Tästä syystä tämä topiikki jätettiin pois myöhemmästä analyysista. 10 seuraavaksi yleisintä topiikkia analysoitiin tarkastelemalla niiden avainsanalistoja ja tviittejä, joissa topiikit eniten esiintyivät sekä tutkimalla topiikkien jakautumista aikajaksoille.

Tulokset

Ennen topiikkianalyysia aloitan aineiston analysoinnin tutkimalla aktiivisimpia tviittaajia sekä yleisimpiä aihetunnisteita. Kaikkiaan aineistossa on tviittejä 42 866 käyttäjältä. 44,1 % käyttäjistä kirjoitti vain yhden tviitin aineistoon. 17 käyttäjää lähetti yli 1 000 tviittiä. Aktiivisimpiin käyttäjiin kuuluu yksityishenkilöitä, uutismedioita, poliitikkoja, botteja ja asiantuntijoita. Aihetunnisteita esiintyy aineistossa runsaasti: lähes 84 % tviiteistä sisältää ainakin yhden aihetunnisteen. Yleisimmät aihetunnisteet liittyvät suoraan koronaan. Muita suosittuja aihetunnisteita ovat hallitus, talous, politiikka ja thl.

Kymmenen yleisintä topiikkia ovat järjestyksessä yleisimmästä vähemmän yleiseen rajoitukset, karanteeni, kiistat ja vieraat maat, rajatarkastukset, kuoleman- ja sairastumisenpelko, julkiset tilat ja palvelut, pandemian leviäminen ja kehittyminen, testit ja oireet, työelämä, ja hallituksen tiedotustilaisuudet. Suurimmassa osassa topiikkeja ei tapahtunut suurta esiintymisvaihtelua aikajaksoittain. Yleisimpiin topiikkeihin liittyvissä tviiteissä korostuvat erityisesti narratiivit käyttäjien henkilökohtaisista kokemuksista koronarajoitteiden kanssa elämisestä. Korona itsessään esiintyy lähinnä taustatekijänä. Lisäksi aineistossa esiintyy vastakkainasettelua rajoitteita noudattavien ja niistä piittaamattomien välillä. Mediassa suurta huomiota saaneet tapahtumat kuten Uudenmaan sulku ja EM-kisaturistien mukanaan tuoma deltavariantti näkyvät myös aineistossa. Kun viesteissä puhutaan koronasta tautina, puhutaan usein sen vakavuudesta, rajoitteiden oikeasuhtaisuudesta sekä ennustetaan tulevaisuuden näkymiä.

Yhteenveto

Tässä tutkielmassa olen analysoinut Twitterissä syntyneitä kriisinarratiiveja koronakriisin aikana. Hyödynsin tutkimuksessa topiikkimallinnusta, jonka avulla tunnistin aineistosta yleisimmät topiikit ja niiden jakautumisen aikajaksoihin. Topiikkehin liittyvissä tviiteissä korostuivat tarinat rajoitusten muuttamasta arjesta, turhautumisesta muiden ihmisten käytökseen, sekä huoli ja murhe taudin uhasta ja leviämisestä. Tämä tutkielma osoittaa, että kriisinarratiivit eivät välttämättä suoraan liity kriisin aiheuttajaan, vaan keskittyvät niihin asioihin, joita ihmiset itse kokevat arjessaan. Tämän ymmärtäminen on tärkeää tulevaisuuden kriiseihin varautuessa.

Topiikkimallinnus on toimiva metodi suurten aineistojen diskurssin tutkimiseen. Kehitystä tarvitaan vielä koherenssimittojen laadun varmistamiseen. Jatkossa tutkimukset samalla aineistolla voisivat syventyä aktiivisimpien käyttäjien tviitteihin ja niissä kerrottuihin narratiiveihin tai tutkia sosiaalisen median ja perinteisen median vuorovaikutusta.