Article

Simon Vydra* and Jaroslaw Kantorowicz

Tracing Policy-relevant Information in Social Media: The Case of Twitter before and during the COVID-19 Crisis

https://doi.org/10.1515/spp-2020-0013
Received November 15, 2020; accepted March 7, 2021

Abstract: Real-time social media data hold great conceptual promise for research and policymaking, but also face substantial limitations and shortcomings inherent to processing re-purposed data in near-real-time. This paper aims to fill two research gaps important for understanding utility of real-time social media data for policymaking: What policy-relevant information is contained in this data and whether this information changes in periods of abrupt social, economic, and policy change. To do so, this paper focuses on two salient policy areas heavily affected by the lockdown policies responding to the 2020 COVID-19 crisis – early childhood education and care policies, and labor market policies focused on (un)employment. We utilize Twitter data for a four-month period during the first wave of COVID-19 and data for the same four-month period the preceding year. We analyze this data using a novel method combining structural topic models and latent semantic scaling, which allows us to summarize the data in detail and to test for change of content between the period of 'normalcy' and period of 'crisis'. With regards to the first research gap, we show that there is policy-relevant information in Twitter data, but that the majority of our data is of limited relevance, and that the data that is relevant present some challenges and limitations. With regards to the second research gap, we successfully quantify the change in relevant information between periods of 'normalcy' and 'crisis'. We also comment on the practicality and advantages of our approach for leveraging micro-blogging data in near real-time.

***Corresponding author: Simon Vydra**, Multi Actor Systems, Delft University of Technology, Delft, Netherlands; and Institute of Security and Global Affairs and Department of Economics, Leiden University, Leiden, Netherlands, E-mail: s.vydra@fgga.leidenuniv.nl
Jaroslaw Kantorowicz, Institute of Security and Global Affairs and Department of Economics, Leiden University, Leiden, Netherlands, E-mail: j.j.kantorowicz@fgga.leidenuniv.nl. https://orcid.org/0000-0002-1186-5427

Keywords: big data analysis, social media data, COVID-19, labour market, early childhood education and care

1 Introduction

The velocity of data is a concept receiving a substantial amount of traction in academia, in no small part due to it being one of the characteristics distinguishing 'big data' from more conventional data sources (Emmanuel and Stanier 2016; Ward and Barker 2013; Ylijoki and Porras 2016). Regardless of the 'big data' concept, a growing part of the data used in research and policymaking is re-used data created by users interacting with various online services, such as search engines or social media, resulting in data streams that are 'always on' (Salganik 2018). Such data streams hold an enormous promise for policymaking, including providing existing indicators faster (Antenucci et al. 2014; Biorci et al. 2017; di Bella, Leporatti, and Maggino 2018), providing them for an area they are currently not available for (Cavallo, Cavallo, and Rigobon 2014; Jean et al. 2016), adding 'depth' or 'detail' to existing indicators (Baker and Fradkin 2011; Proserpio, Counts, and Jain 2016; United Nations 2011), or even creating novel indicators (di Bella, Leporatti, and Maggino 2018; Turrell et al. 2019). However, there are also great challenges related to the comparatively little control researchers and policymakers have over the data generation process and to the difficulty of integrating processes to guarantee reliability, accuracy, or representativeness of data into a 'near real-time' timeframe.

Nowhere is this tension between grand promises and substantial challenges more apparent than in the case of social media data: Conceptually, this data is a gold-mine of information about important events in peoples' lives and their perception of them. In the case of Twitter, the social media platform this paper focuses on, the two primary motivations for using the platform are to connect with others and to seek or share information and advice (Chen 2011; Johnson and Yang 2009), which results in users posting updates about their life (Java et al. 2007) or sharing their beliefs and concerns with regards to current (crisis) events (Gilardi et al. Forthcoming; McNeill, Harris, and Briggs 2016; Signorini, Segre, and Polgreen 2011). Such tweets provide very detailed information at micro-level and in near real-time, lending them well to being utilized in the policymaking process, especially in situations requiring a rapid policy response. However, in the area of social and economic policymaking the existing research that utilizes social media mainly focuses on replicating conventional indicators like unemployment (Antenucci et al. 2014; Biorci et al. 2017; Proserpio, Counts, and Jain 2016) and, despite sometimes providing more 'depth' to these indicators

(Proserpio, Counts, and Jain 2016), the more ambitious theoretical promise of novel real-time indicators and policymaking tools remains under-explored. This constitutes the first research gap this paper explores – the lack of knowledge with regards to what relevant information for social and economic policymaking do real-time social media data streams contain.

The second research gap this paper explores is whether the relevant information in real-time social media data changes in periods of abrupt social, economic, and policy change. This is a research gap the literature pays little attention to despite the fact that real-time social media data (and Twitter specifically) have been used to map and understand public reaction to crises ranging from short-term crises like natural disasters (Acar and Muraki 2011; Terpstra et al. 2012) or mass shootings (Heverin and Zach 2010) to longer term crises like the refuggee crisis (Gualda and Rebollo 2016; Öztürk and Ayvaz 2018). These cases include the 2009 H1N1 outbreak (Ahmed et al. 2019; Chew and Eysenbach 2010; McNeill, Harris, and Briggs 2016; Signorini, Segre, and Polgreen 2011; Szomszor, Kostkova, and De Quincey 2011) as well as the COVID-19 outbreak (Gilardi et al. Forthcoming). We thus know that people use social media like Twitter to share information about the crisis and their personal experience with it, but we lack the understanding of whether there is a meaningful change in what they say about social and economic policies that are not established as a result of the crisis, but that are affected by it. Intuitively we can hypothesize some degree of change, but the empirical work testing such a hypothesis is currently not available – existing research focuses on direct response and reaction to a crisis event (which simply did not exist prior to the crisis event) making any comparison with the pre-crisis period trivial and not informative.

Understanding this change (or lack thereof) is important for our understanding of the value of big data in policymaking: The key advantage of real-time data streams is the ability to support a faster policy response, but there is little incentive to make fast decisions and abruptly change a policy suite in a time period of normalcy and stability. This can result in very little incentive to trade-off the generally higher accuracy and reliability of data sources like household samples surveys for the higher velocity of real-time data in dealing with some policy puzzles (Vydra and Klievink 2019). In other words, the potential benefits of using real-time data to support policymaking are greatest in a time of crisis, such as the 2020 COVID-19 outbreak, when policy needs to be changed in a timeframe in which traditional data will not be available. Yet, we have little understanding of how the content of real-time social media data changes in such situations with respect to existing policies. It could be that affected policies get debated even more (since they are more salient) or that the crisis itself creates new grievances that people comment on. It could also be that the content is too narrowly focused on the

crisis event itself or that generic and politicised 'noise' drowns out the 'signal' of meaningful commentary. Having an empirical understanding of the change that happens can inform our understanding of the transformative potential of real-time data for social and economic policymaking.

This paper attempts to (partially) fill both of these research gaps by empirically studying tweets (as an instance of real-time social media data) focused on (un) employment policies and early childhood education and care policies – two well-established policy areas heavily affected by the COVID-19 outbreak. We utilize data for two time periods – a four-month period of 'crisis' following the 2020 COVID-19 pandemic, and the same time period from 2019 as a period of 'normalcy'. This allows us to tackle two research questions: Firstly, the more exploratory question **'What policy-relevant information does Twitter contain?'** This broad question necessarily touches on whether the information exists, what insight it carries, how much of it there is, and how well are we able to extract it. Secondly, the more descriptive question **'How does this information change between a period of normalcy and a period of crisis?'** By 'policy-relevance' we refer to information where individuals express their opinion on the (in)sufficiency of relevant policies, specific aspects of those policies, or the situations those policies aim to address (e.g. joblessness). Our approach is distinct from other approaches utilizing social media data to measure policy-relevant indicators (Proserpio, Counts, and Jain 2016) or issues associated with the COVID-19 pandemic (Gilardi et al. Forthcoming) in having an a-priori determined (and rather broad) focus and not relying on sampling individual users. This approach can result in somewhat general information, but this information is not meant to replace a thorough policy evaluation in practice and in terms of testing the theoretical promise of real-time data for policymaking we have to avoid an overly restrictive focus to maintain any ability to generalize.

We answer the two research questions for the case of the Netherlands during the 2020 COVID-19 outbreak, focusing on two policy areas to improve the external validity of our findings. We select these policy areas due to their general political salience, but also due to how severely they were impacted by the lockdown policies responding to the COVID-19 pandemic. We describe our case, policy, platform, timeframe selection, and conception of 'policy-relevance' in detail in Section 2. We then analyze this data using a novel methodological approach that combines topic modelling and latent semantic scaling – an approach that we propose and justify as fitting for this particular paper in Section 3. In Section 4 the results are presented, and we discuss the policy-relevance of our findings. Section 5 mentions limitations of this paper and Section 6 concludes by answering the two research questions and commenting on our approach as a whole.

2 Case and Data Selection

To meaningfully test for presence of policy-relevant information, we constrain our dataset to a single country and to two relevant policy areas. This section justifies our country selection – the Netherlands – based on high take-up rates and salience of the two selected policy areas (Section 2.1) as well as social media utilization. It further substantiates the policy selection of early childhood education and care (ECEC) policies and (un)employment policies based on their political salience and the degree to which they are affected by the COVID-19 crisis (Section 2.2). It then argues (Section 2.3) for utilizing Twitter as a social media platform due to its fit with the two selected policy areas as well as pragmatic reasons and for the two four-months data collection periods (Section 2.4). Most importantly, we define what constitutes a 'policy-relevant' piece of information (Section 2.5), adopting a crowdsourcing approach to creation of socioeconomic indicators.

2.1 Country Selection

This paper focuses on the Netherland for two primary reasons: Firstly, from a practical perspective the most important factors are a very high Internet penetration which the Netherlands has, together with high rate of social media use (Internet World Stats 2017), which is necessary to gather a sufficient amount of data representing a broad section of the population. Secondly, both of the policy areas selected for this paper (unemployment and childcare policy) are well developed in the Netherlands: The Netherlands has a relatively low percentage of children under the age of 3 who receive no formal childcare (35.2% for 2019) (Eurostat 2020). In 2018 the Netherlands was the third lowest in Europe in this statistic followed only by Luxembourg and Denmark (Eurostat 2020). Multiple childcare options are also comprehensively supported by the government. In terms of unemployment policies, the Dutch expenditure on out-of-work income maintenance and support is the third highest in Europe (1.38% of GDP in 2018) (European Commission 2020), but the contributory conditions are some of the strictest in Europe (Matsaganis, Ozdemir, and Ward 2014). Data on unemployment assistance take-up rates is not available for the Netherlands.

2.2 Policy Selection

There are two groups of policies that this paper focuses on – labour market (LM) policies related to (un)employment as well as early childhood education and care

(ECEC) policies. Tackling two policy areas rather than one is simply to improve the external validity, as it is to be expected that the public assessment and overall conversation will be different for different policies. There are two main reasons for selecting these policies: Firstly, it is because of the general political salience of these two policy areas. On a European level, employment is an explicit goal of the Europe 2020 agenda, and childcare contributes to both the employment (of the parents) as well as the educational goals. This translates well to Dutch policy priorities – as mentioned, the Dutch utilize childcare a lot and unemployment policies are sufficiently generous. Secondly, and more importantly, it is because of how heavily impacted these policy areas are by the COVID-19 outbreak. The amount of global working hours has decreased by approx. 17.3% in second quarter of 2020 (compared to Q4 2019) with women affected more severely, in part due to the increased burden of unpaid labour (ILO 2020) such as caring for children. In the Netherlands, unemployment has increased from 2.9% in March 2020 to 4.6% in August 2020 (CBS 2020), which is despite the government's intervention providing companies with financial support to pay their employees and providing financial support, credit, and relaxed taxes to the self-employed. In terms of childcare (and schools), the access was restricted for everyone except children of crucial workers in late March, and the restriction was lifted on May 11th. No official statistics are available on how this impacted the number of children formally enrolled in childcare during the restrictions or immediately after. Both labour markets and childcare options have been heavily affected globally, and the Netherlands is no exception to this. Such a period of abrupt change to the labour markets and childcare is a well-fitting instance of a 'crisis' situation, where substantial policy intervention is required, but where traditional data is not available in time and will not provide sufficient detail.

2.3 Platform Selection

For data collection Twitter is the platform of choice for several reasons. Firstly, the demographics of Twitter are a good fit with the policies we focus on. In the Netherlands the best represented demographic group on Twitter are those between 20 and 39 years of age (Statista 2018), which is the prime age for activity in the labour market as well as starting a family and child rearing. Furthermore, it is also reasonable to expect that these two policy areas will in some way be debated on social media. In terms of childcare, new mothers often seek support from their networks on social media (McDaniel, Coyne, and Holmes 2012) and sharing ones experiences or looking for advice about childcare options is also something we hypothesize to see in social media data. With regards to employment, one's

personal network is an important tool that can be leveraged using social media platforms. In other words, we select policy areas that we expect to be discussed on social media more than other policies, and we select policy areas likely to be of interest to the demographic group most heavily represented on social media (in this case on Twitter). Despite other platforms such as Facebook having a substantially higher user-base in the Netherlands, Twitter maintains 2.5 million users in the Netherlands for 2019 (Statista 2019) and a 2018 poll reports 26% of 20–39 year olds in the Netherlands using Twitter (Statista 2018), which is good comparatively to other social media platforms. Furthermore, as much as the debate on Twitter is influenced by 'opinion leaders', opinion leaders on Twitter do not necessarily consume a lot of traditional media (Park 2013) and do not share the same socio-economic characteristics of 'offline' opinion leaders (Park and Kaye 2017). This suggests that the discussion on Twitter is unlikely to just be a reflection of the narratives found in traditional media and thus it should be capable of providing additional insight.

Choosing Twitter is also a pragmatic choice due to the accessibility of research data compared to other social media platforms where this access is not provided and automatically scraping the platform would violate the terms of service as well as users' expectations with regards to privacy. Secondly, this data is by default available in a real-time data stream, which is not the case for other social media data that would either have to be retroactively scraped or retroactively searched for.

2.4 Timeframe Selection

This research gathers Twitter data for two time periods – from the 11th of May until the 11th of September 2019 and 2020. This selection is of course limited by practical constraints, but the 11th of May starting point is selected to coincide with the re-opening of schools and day-care centres (to non-essential workers) during the COVID-19 outbreak. With the re-opening, these facilities can be used again but people will likely have concerns that are not captured in any conventional data source. This re-opening also means that a lot of parents could focus on their labour market situation with the children being back to school/day-care. Understanding these developments in near real-time could be crucial for designating appropriate policy action. The end point for our data collection assures we capture the 'first' wave of the COVID crisis and the start of a new academic year. With regards to unemployment assistance the start and end date are not as consequential, but the time period in 2020 covers the period of abrupt unemployment increases as well as the availability of special governmental assistance. The same four-month period

from 2019 is used as a period of policy 'normalcy', minimizing the effect of seasonality.

2.5 Defining 'Policy-relevance'

The most consequential definition adopted by this paper is what it means for information on social media platforms to be 'policy-relevant'. This is due to the fact that any such definition carries with it assumptions about the functioning of policymaking as well as about the specific data needs in policymaking decisions.

Starting at the broadest possible level of the general role of social media data in governance: The extant literature outlines a range of potential uses like electronic participation, engagement, transparency, communication, trust, collaboration, democracy, crowdsourcing, security, and open data practices (Dwivedi et al. 2017). In practice, the ways governments utilize social-media are also quite varied, ranging from utilization in elections, information dissemination, making processes transparent, but also sourcing information and feedback from users to be utilized in decision-making (Grubmüller, Götsch, and Krieger 2013). In this paper we focus exclusively on using social media as a 'information and feedback source' (Grubmüller, Götsch, and Krieger 2013) for decision-making, which is also known as crowdsourcing. Crowdsourcing involves problem-solving, idea-generation, and production tasks that use IT to leverage the dispersed knowledge of a 'crowd' of individuals (Prpić, Taeihagh, and Melton 2015). This knowledge is then utilized as evidence in evidence-based policymaking.

With our focus on crowdsourcing set, the next question is how should the 'evidence' we are seeking to find be utilized, as there is no one-size-fits-all answer with regards to how evidence from social media enters existing policymaking practice (Höchtl, Parycek, and Schöllhammer 2016; Janssen and Helbig 2018; Mergel and Bretschneider 2013). For the purposes of this paper we assume that the information we aim to find is relevant mainly for the latest and the earliest stages of the policy cycle: Either the information can be used as near-real time monitoring tool to evaluate the perceptions of and actual problems with a specific policy (Singh et al. 2020), or it can be leveraged in an exploratory way during agenda setting and problem definition (Panagiotopoulos, Bowen, and Brooker 2017). This makes our focus somewhat broad as we are not only looking for very policy-specific commentary in which people voice concerns about specific policy (or its aspects), but also for more general commentary where people identify problems with the situation those policies aim to address (childcare situation and employment situation). We do this for two reasons: Firstly, to avoid excluding information where users have genuine and relevant complains but simply do not

link them to a policy explicitly. Secondly, to allow policymakers and researchers a degree of freedom in determining what policy can solve a given issue – even if people link a problem they are experiencing to a policy they might not do so 'correctly', especially given that many problems can be solved by different policies.

Given our policy selection and focus on crowdsourcing information to be used as a monitoring and/or agenda setting tool, our empirical contribution is most closely aligned with literature on real-time indicators for economic policymaking. In this area, indicators based on real-time social media data have a huge advantage over traditional household and business survey approaches (Antenucci et al. 2014): First, (labour market) indicators created in real-time allow for a more rapid diagnosis of an issue and a timely policy response, especially when it is most appropriate such as in times of crisis. Second, they can enable a more targeted policy response by, for instance, identifying socio-economic or geographical groups most hit by a crisis, and by pointing to particular aspects of policy which demand adaptations. Third, this type of data comes at relatively low cost as this information exists despite its potential utility for the policy insight. The costs stem from creating an appropriate system of retrieving relevant information and maintaining it later on; thus at arguably lower cost than running consumer surveys. These benefits have already motivated efforts to create real-time social media indicators with varied degrees of success (Antenucci et al. 2014; Biorci et al. 2017; Proserpio, Counts, and Jain 2016; United Nations 2011). In terms of adding additional 'depth' to economic indicators, these efforts range from simple replication of the unemployment indicator (Antenucci et al. 2014) to understanding psychological impacts of unemployment (Proserpio, Counts, and Jain 2016) or coping mechanism (United Nations 2011). There is currently no research (that we would be aware of) testing the merit of such indicators in periods of crisis, but in terms of other real-time data sources, there is some promise for tracking economic activity in response to the COVID-19 crisis. A good example is research by Chetty et al. (2020) which was able to quickly identify relative ineffectiveness of state-ordered reopening to stimulate economic activity in the aftermath of the lockdown and argued that the only effective approach to mitigating the short-term economic hardship is through providing benefits to those who have lost their incomes (Chetty et al. 2020). It thus offered a clear recommendation for policy adaptation. While this research was based on economic transaction data provided by private entities, we test in this paper Twitter's utility in providing relevant real-time information which could trigger policy adaptation, especially in periods of crisis.

Conceptually restricting 'policy relevance' to **crowdsourcing** of **real-time socioeconomic indicators** to inform the **monitoring and/or agenda setting**

stages of policymaking is necessary to firmly situate this paper in the existing literature, but it does not solve the practical issue of differentiating between 'relevant' and 'irrelevant' tweets. To do that, we adopt a simple and intuitive definition: 'Policy relevant' tweets comment on the adequacy or inadequacy of specific policies, specific aspects of those policies, or the situations relevant policies aim to address. In the analysis we operationalize 'adequacy/inadequacy' in multiple ways, but in general we are interested in people identifying issues that make them (un)happy with a policy and that render the policy somehow (in) effective or (un)desirable. What we mean by 'aspects of policy' here are different features of a policy provision as perceived by individuals; For unemployment policy these could include the generosity, length, or eligibility criteria or unemployment benefits (Gallego and Marx 2017). For childcare these could include affordability, capacity, or quality of care (Carta and Rizzica 2016; Grammatikopoulos et al. 2014; Kawabata 2012). Trying to differentiate between these individual aspects in both policy areas is important because negative comments on one such aspect of a policy convey fundamentally different feedback than those on another.

Despite defining policy relevance and the practical rules making it applicable to individual tweets, our decisions about which aspects of policies are relevant are ultimately subjective and in practice would be made by policymakers. In the absence of a real-world policy dilemma and extensive cooperation with policymakers responsible for resolving it, we can only approximate 'policy relevance' by assuming that factors that are either shown (in the literature) to influence the effectiveness of a policy, or that we hypothesize to influence the effectiveness of a policy are 'policy relevant'. For the purposes of this paper this is not a limitation per se, but we do address this issue in articulating our methodology (section 4) by introducing specific steps where policy relevance is decided without necessitating a technical understanding of other stages of the method, making it relatively easy to include policymakers at appropriate points in the analysis.

3 Methods

To answer the research questions of this paper we propose an approach combining topic modelling and latent semantic scaling (LSS), both of which are machine learning methods developed for automated text analysis. In this approach we first establish (using topic modelling) the substantive focus of individual tweets allowing us to extract clusters of tweets (topics) focusing on relevant policies or situations those policies aim to address. However, such a

summary does not always tell us which policy aspects are being talked about or what exactly is being said about them as the topics can focus on multiple aspects of a policy or convey multiple opinions about a policy. Topic modelling alone does not provide this insight, which is why LSS is utilized to further summarize the content in these topics along policy-relevant 'dimensions' that we specify. These dimensions can be rather general such as positive sentiment versus negative sentiment (the main intended use of LSS), but also more tailored to policy such as succeeding versus failing. The two methods are used in sequence with topic modelling first identifying a group of tweets relevant to a topic of interest and LSS models then estimating whether these tweets tend to be negative or positive and concerned with success or concerned with failure (as those are the two LSS dimensions we construct). The LSS scores for groups of tweets that we report in the results section and in appendix B are simply the mean polarity score of the selected group of tweets, which also allows us to run a statistical test for the difference in mean/median between tweets posted in a period of normalcy and tweets posted in a period of crisis.

The general methodological approach and how it answers the research (sub) questions is illustrated in Figure 1. This figure outlines what we consider major steps in the analysis process, with dashed borders denoting the two steps where the meaning of 'policy-relevance' is established. In practice many analytical decisions made along the way can carry subjective assumptions, but the two highlighted steps are almost purely subjective and normative and when applied to real-world policymaking would necessitate the involvement of policymakers.

This approach is able to answer the first (more exploratory) research question by identifying relevant topics in the corpus and then testing whether tweets representative of those topics have a polarity score significantly different from zero on LSS dimensions. It then builds on this finding in order to answer the second (more descriptive) research question by utilizing two analytical outputs: Firstly,

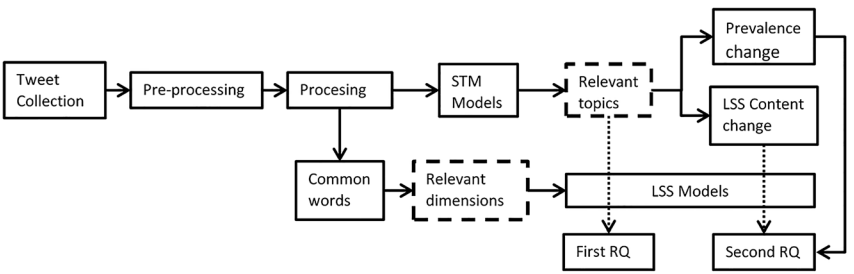


Figure 1: Method diagram.

whether there is a significant difference in the prevalence of policy relevant topics between a time period of abrupt change and a time period of normalcy. Secondly, whether there is a significant change in LSS polarities between these two time periods for a given topic. This allows us to answer both research questions of the paper not just at a level of detail unparalleled by other methods but also by using statistical tests of difference, providing a more definitive conclusion than one based on our qualitative interpretation of topics and their change.

Outside of fitting the research focus of this paper, this approach has multiple important and more general advantages: Firstly, it can leverage the real-time nature of social media data by providing outputs in near-real-time. There are steps of the process that cannot happen in near-real-time, such as training the full models or the two more normative steps that determine 'policy-relevance' (highlighted by dashed borders in Figure 1), but once those steps are taken, the models we utilize can take new tweets and predict their topic memberships and their polarity on relevant dimensions in near-real time. These underlying models will need to be iteratively updated to avoid them becoming out-dated, but applying them to a stream of tweets can leverage its real-time nature. Secondly, the two instances where the models have to be 'supervised' are actually an advantage: As mentioned earlier, decisions on which information is 'policy-relevant' are inherently subjective and to be determined by policymakers rather than technical experts. The fact that these decisions are transparent and can be made in a very intuitive way without requiring technical expertise allows policymakers to be involved, make informed decisions, and be held accountable. Thirdly, both topic modelling and LSS are entirely language-independent, making the entire approach language-independent as well. Processing the Tweets themselves can be language-dependent, in our case it is, but there are also language-independent options and any potential language-dependence is easier to overcome because it is limited purely to the 'Processing' step. Fourthly, this approach is not specific to Twitter as a platform and can be easily adapted to other user-generated text data such as other social media or comments on governmental platforms. Fifthly, it allows for good internal validation by validating the STM models themselves, the LSS models themselves, as well as the final summary by inspecting tweets and assessing whether a topic theta (belonging to a topic) and LSS polarity estimate (positive/negative or success/failure) are assigned in a way that is interpretable and agreeable to human annotators. We cover the validation of STM and LSS models themselves in this section and comment on the validity of the final output in a topic-by-topic fashion in appendix B.

The rest of this section describes the methodological steps in more technical detail in three sub-sections focusing on the collecting, pre-processing, and

processing of the data (Sub-section 4.1), on the topic models used and their parameterization (Sub-section 4.2), and on LSS and creation of relevant 'dimensions' (Sub-section 4.3).

3.1 Collection, Processing, and Pre-processing

The data was collected via Twitter's Stream API by gathering tweets that are a) in the Dutch language (as identified by Twitter) and b) containing some of the keywords from a list of keywords aimed to capture relevant labour market and ECEC policies, as well as situations those policies aim to address (list of keywords available in appendix A). Pre-processing includes removing re-tweets and duplicate tweets and joining quoted tweets with the text of the tweet quoting them (including '|' between the two texts). Bot removal is very rudimentary and consists of removing tweets authored by accounts tweeting more frequently than 1500 times a month and/or authoring more than 450 tweets a month that get captured in our dataset. This issue is likely better tackled as a supervised learning problem (Andriotis and Takasu 2019; Inuwa-Dutse, Liptrott, and Korkontzelos 2018; Kantepe and Ganiz 2017; Lee, Eoff, and Caverlee 2011), but such an approach necessitates ground-truths about which accounts are bot accounts and would sacrifice the language-independence and easy applicability of the general method we are proposing. Given that we focus our interpretation on specific topics, the existence of other topics that include some content from bot accounts is not a major drawback (any remaining irrelevant content authored by bots likely gets removed by omission at the stage of selecting relevant topics).

The entire resulting dataset (approx. 740,000 tweets) is used in training LSS models, but for training the topic models we create two corpus sub-sections, one concerned with ECEC (approx. 40,000 tweets) and one with labour markets (approx. 403,000 tweets). We do that utilizing groups of the same keywords used to collect the data. Tweets that are not included in either corpus are excluded because not all keywords used for collection are used to create corpus sub-sections (see appendix A for details on keywords). Text for all tweets is then tokenized and lemmatized – splitting it into individual tokens that get reduced to the 'lemma' of a word, which is its basic dictionary form. This resolves the issue of words that have inflectional and derivationally related forms appearing as different tokens despite being (effectively) the same word. To do this step we utilize the 'Frog' natural language processing suite (Van Den Bosch et al. 2007).

3.2 Topic Models and Parameterization

Topic models are a class of algorithms that aim to identify existing 'themes' of a corpus of documents and can organize the documents according to those themes at a scale where human annotation would be near-impossible (Blei 2012). These models operate on the assumption that documents can be summarized as collections of various 'topics' and that those topics can be summarized as collections of various tokens. These assumptions are most famously formalized into a generative probabilistic model called Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), which has since become widely adopted and utilized for its simplicity and good interpretability. Since then topic models have expanded to take into account correlation between topics (Blei and Lafferty 2007), the evolution of topics across time (Blei and Lafferty 2006), the evolution of topics due to influential documents (Gerrish and Blei 2010), or to be able to train in real-time from a data stream (Wang, Agichtein, and Benzi 2012). The model utilized in this research is the Structural Topic Model (STM) (Roberts, Stewart, and Airoldi 2016) and its R implementation (Roberts, Stewart, and Tingley 2019). This model includes the information about topic correlations but more importantly assumes that topic prevalence and/or topic content can be influence by a generalized linear model of document-level covariates.

The ability to include document-level covariates into the analysis makes STM the best fit for this paper for two reasons: Firstly, Twitter is a platform that different users use in different ways and we would expect the accounts of political parties, companies, or individuals to address different topics. STM allows us to include document-level covariates describing the type of account posting a Tweet (such as follower count or tweeting frequency) as factors that influence either the prevalence or the content of topics (or both). This makes the model more conceptually fitting than models that would force us to 'assume away' the difference between users. Secondly, it allows us to answer the research sub-section about how relevant information changes in times of abrupt social, economic, and policy change by simply including a dummy covariate capturing whether a tweet was posted in a period of abrupt change or a period of normalcy. With regards to assuming change in topical prevalence and/or topical content, our baseline is to assume change only in topical prevalence as that makes topics easier to understand and provides us with more solid footing to run a statistical test of difference on mean or median polarity (if we assume a change in the topic itself between those two periods we risk comparing 'apples to oranges'). We thus assume a simple generalized linear model for prevalence change of topics: Prevalence = tweeting frequency + number of followers + crisis or normalcy. To test this difference in polarity we adopt a

two-sample *t*-test and a Mann-Whitney *U* test: A two-sample *t*-test is generally sufficient due to the size of the data, but the polarity distribution is often bi-modal (due to the fact that we are looking for non-zero polarity), data can be small in size, and samples can be imbalanced. A potential divergence between the *t*-test and Mann-Whitney prompts us to investigate a test closer. We report the results of both tests in the results section.

With any topic model there is the challenge of determining an appropriate number of topics to model for. There is not a 'wrong' topic number in topic modelling (Grimmer and Stewart 2013; Roberts, Stewart, and Tingley 2019) as the quality of topic models generally depends on their interpretability (to humans) and the insight they deliver. The quality of the insight depends entirely on why the model is used and interpretability is highly subjective and dependant on domain-specific knowledge about a corpus. That said, there are multiple metrics assessing the quality of models that can aid with the selection of a topic number. These metrics are presented for both the ECEC corpus sub-section and the LM corpus sub-section in Figures 2 and 3 respectively and there are four of them:

1. Semantic coherence: This is a per-topic metric and is maximized when the top words of a topic tend to co-occur in the corpus. This metric is important because it correlates with human judgment of topic interpretability (Mimno et al. 2011). However, Roberts et al. (2014) show that this metric can be maximized by having topics dominated by relatively common words resulting in semantic coherence generally declining as number of topics increases making coherence insufficient on its own.
2. Exclusivity: This is a per-topic metric that remedies the problem of coherence by providing a measure of how exclusive assigned words are to a given topic, scoring models where general words are shared across many topics poorly. This metric should be read together with semantic coherence to select topic numbers where both metrics are relatively high (but trade-offs are unavoidable).
3. Heldout likelihood: This is a metric to be maximized and is calculated by removing a proportion of words (in our case 40%) from a proportion of documents (in our case 10%) before training the model and then measuring how well the model predicts the missing words.
4. Dispersion of residuals: Metric utilized by Taddy (2012) computes the dispersion of residuals and should be minimized as over-dispersed residuals can mean that the variance has not been sufficiently accounted for.

We use these four metrics to determine a reasonable range of topic numbers, but the final selection is made based on manual inspection of models in that range. This is because exclusivity and semantic coherence are computed per-topic and plotting the averages for a model discards information about variance in this

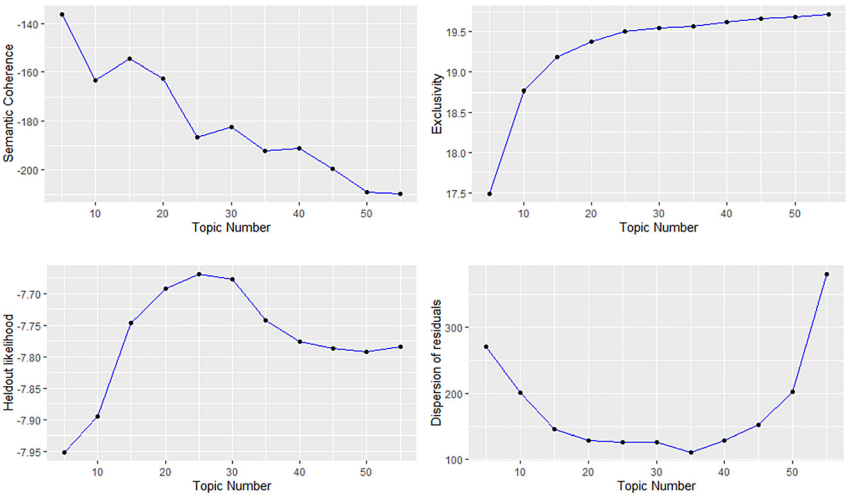
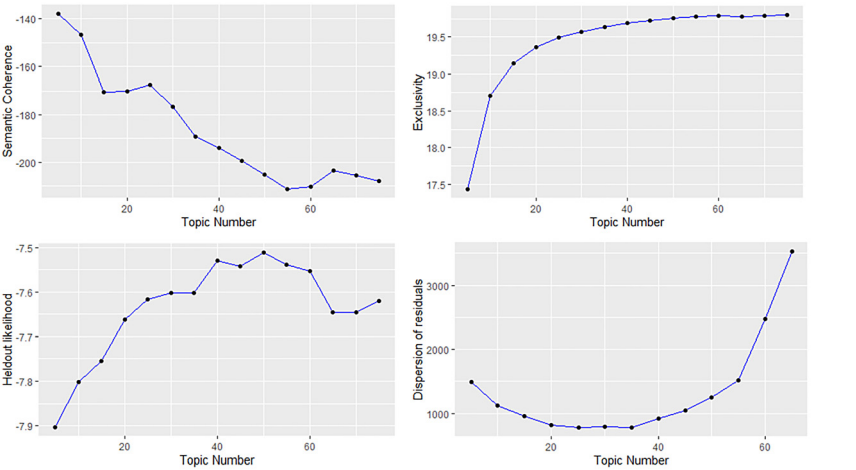


Figure 2: Model metrics for ECEC sub-section.



Note: For the dispersion of residuals models with 70 and 75 topics are excluded as they are outliers and including them stretches the scale thus obscuring the trend.

Figure 3: Model metrics for LM sub-section.

metric. Figure 4 illustrates this on the candidate models: 20–30 topics for the ECEC sub-section and 25–40 topics for the LM sub-section. From Figure 4 we observe that each model clearly contains topics that are relatively coherent and exclusive, meaning that any of these models can provide good insight given that we will be focusing only on 'relevant' topics and discounting the rest. We inspect models by first looking at 15 'top words' for each topic in each candidate model and for seemingly interpretable topics we inspect 30 tweets – six groups of five tweets along a range of theta values (maximal, 0.85, 0.8, 0.75, 0.7, 0.65), which represents documents that are 'representative' of a topic to a given degree. We do this to get a deeper insight into the topic (avoiding issues such as top tweets all belonging to the same conversation), but also to provide a layer of internal validation for our topic models: This inspection allows us to select the most 'valid' model and to also identify which topics validate and which do not, which is simply judging of whether the model assigns tweets to 'correct' topics with a 'correct' theta as assessed by human observers (in this case the authors). Using this method of manual inspection, we select 20 topics for the ECEC sub-section and 30 topics for the LM sub-section and identify topics that are potentially policy-relevant and clearly validate (in the sense of being interpretable by humans).

3.3 Document Scaling Using LSS

The other method we utilize is Latent Semantic Scaling (LSS), which is a semi-supervised technique for document scaling that has been shown to perform comparably to lexicon-based approaches for sentiment analysis (Watanabe 2020) and has been used as a language-independent sentiment analysis method (Watanabe 2017a; b). It relies on a word embedding approach that expresses individual tokens as vectors that can then be compared to assess the semantic

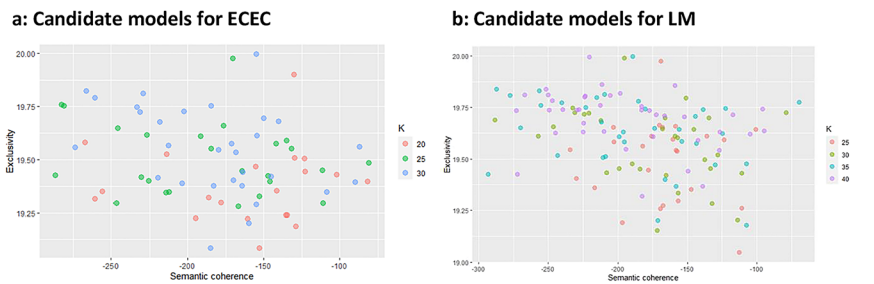


Figure 4: Exclusivity and Semantic coherence across multiple topic models.

similarity between them. LSS relies on singular value decomposition of a document-term matrix representation of a corpus where each document is a singular sentence (Watanabe 2020). Sentences are the context in which words are considered to co-occur for the purposes of training a word embedding model (which is appropriate for our corpus of rather short texts). Researchers are then required to provide seed words that represent two extremes of a dimension of interest and LSS infers the polarity of other word vectors based on their proximity to the two defined extremes. Conceptually these dimensions do not have to be limited to sentiment and can be constructed in various ways – the limits of what such dimensions can meaningfully capture are to still be tested by the literature, including by this paper.

LSS as a method is appropriate for this paper for a multiplicity of reasons. The Dutch language does not have as many options for sentiment analysis as English, and existing lexicon-based approaches do not perform well on our corpus (language on twitter tends to be highly informal and as a result majority of tweets cannot be annotated reliably). Secondly, the ability to control exactly what a dimension captures is crucial for this research, as sentiment alone, despite carrying valuable information, is not perfect: Identifying that tweets about a specific policy tend to be negative could be users expressing displeasure with the policy, but it could also be users condemning the problem that the policy is addressing (and voicing support for intervention). We try to partially resolve this problem by creating a more policy-specific dimension of success (at one extreme) and failure (on the other extreme). This dimension bears some similarity to the sentiment dimension but also limits the issue of negative sentiment being associated with a situation rather than intervetion mentioned above. Many other dimensions of interest can be constructed, but the selection of those is ad hoc without the involvement of policymakers who would articulate a particular interest.

However, the ability to define the relevant dimensions comes with a drawback – these dimensions need to be validated and interpretable (beyond the provided seed words). To train LSS models for individual dimensions we utilize the full corpus of tweets and remove all '@' mentions to remove unwanted bias (without doing so, certain political parties and public figures would be strongly associated with certain extremes, constituting an undue bias). We include a form of internal validation into constructing these dimensions. We base our keyword selection on a manual inspection of top 500 most commonly occurring adjectives, adverbs, verbs, and nouns (assuming that conjunctions, prepositions, etc. do not convey strong polarity). How many times the token appears in the corpus is included as 'Count' in Table 1. We then pick the most relevant words from that list that convey polarity with respect to a given dimension. We provide these tables for the dimension positive sentiment versus negative sentiment in Table 1a and or the dimension success versus failure in table Table 1b. For each dimension we seed only a part of the selected words into the model

and the rest are later displayed along the given dimension to validate whether they are placed 'correctly' as judged by human annotators (in this case the authors). The resulting plots are displayed in Figure 5 (for positive vs. negative sentiment) and 6 (success vs. failure) – the words that were not seeded into the models are those highlighted in grey in Table 1. Rather than randomizing which words are seeded and which are heldout for validation we select them manually to include word pairs that are clearly antonyms ('good' & 'bad' or 'positive' & 'negative') and words that are not ambiguous in terms of polarity. This is especially important for Dutch as many generally positive words can simply be preceded by 'niet' (translates as 'not') to capture the opposite, making those words very ambiguous as they can be used in both negative and positive contexts. In general the sentiment dimension validates well as the heldout words are placed towards the 'correct' polarity to an understandable degree (Figures 5 and 6).

Another form of validation is inspecting the words assigned the strongest polarity in either direction. Doing so for the example dimension validates it further but also reveals that the dimension captures positive and negative sentiment

Table 1: Selected words for LSS Dimensions.

1a: Positive vs. negative sentiment

Positive sentiment			Negative sentiment		
Count	**Token**	**Translation**	**Count**	**Token**	**Translation**
127810	goed	good	9262	slecht	bad
2834	positief	positive	1290	negatief	negative
5490	Blij	happy	1375	boos	angry
4913	gelukkig	happy	968	zielig	pathetic
930	tevreden	satisfied	472	kwaad	pissed off
3203	prachtig	magnificent	616	idioot	idiot
21721	mooi	beautiful	1039	belachelijk	ridiculous
1076	lief	sweet	2972	dom	stupid

1b: Success vs. failure

Success			Failure		
Count	**Token**	**Translation**	**Count**	**Token**	**Translation**
1917	succesvol	successful	634	onnodig	redundant
1766	handig	useful	511	mislukken	fail
2409	lukken	succeed	804	falen	to fail
1495	behalen	achieve	2660	verliezen	to lose
1866	bereiken	to achieve	473	nutteloos	useless
1159	afmaken	finish up	1930	tekort	shortage
1100	realiseren	realize			

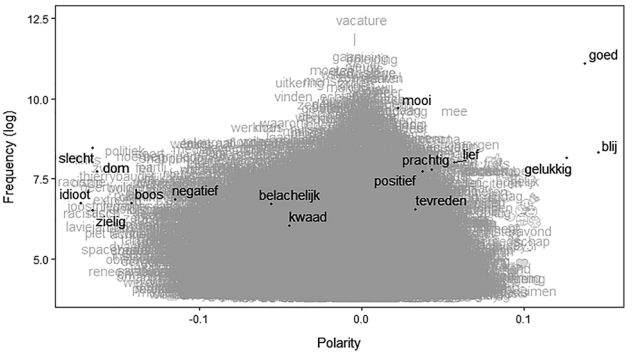


Figure 5: Words on a positive vs. negative sentiment dimension.

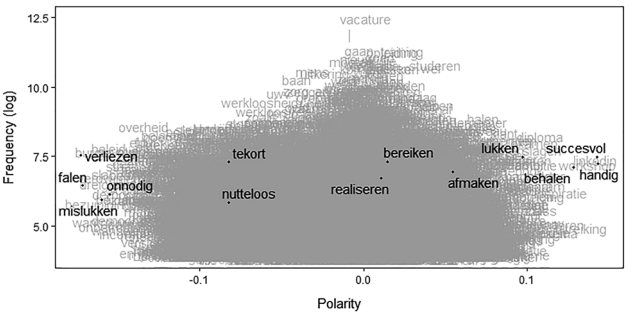


Figure 6: Words on a success vs. failure dimension.

specific to different topics. In this example the most 'positive' words include words like summer, sunny, vacation, and various happy emojis. The most negative words include racist, racism, nonsense, politics, the names of right-wing political parties and commentators, and Black Pete (captured by tokens 'zwart' and 'piet', referring to a controversial Dutch tradition that many consider racist). In other words, the negative extreme is related to political sentiment and the positive extreme is related to sentiment about weather. This is not (necessarily) a wrong representation of the data, but might not be desired for certain topics. LSS includes an approach that resolves this issue by restricting the model terms to only terms that tend to co-occur with selected tokens (effectively restricting model terms to a given theme or topic), but we do not always find meaningful improvements in utilizing it and it sacrifices the generality of our dimensions. Furthermore, our approach utilizes STM modelling to identify tweets representative of a certain topic prior to applying and interpreting LSS estimates. This means that even if the dimension is

somewhat general it is applied to a rather specific collection of tweets, reducing the risk of providing a misleading summary.

The last and most important form of validation comes from predicting a polarity score of individual tweets, which is done for every dimension applied to every topic of interest (manually inspecting 20 tweets with the highest and the lowest polarities that 'contain' at least 50% of a given topic). This is important because even a dimension that seems interpretable might not be interpretable when applied to tweets from a specific topic, either due to technical reasons or due to the nature of the topic cluster.

4 Results

Applying the proposed method we focus on a 20-topic model for the ECEC sub-section and a 30-topic model for the LM sub-section of our corpus. These models generally do not deliver a fundamentally different insight to other candidate models, but they tend to deliver more 'focused' topics. We present summaries of the full models in Appendix B. From these models we mainly focus on four topics from the ECEC sub-section and five topics from the LM sub-section as those topics interpretable and potentially policy-relevant: From the ECEC sub-section we focus on **topic 5** concerned with ECEC benefits and their beneficiaries, **topic 11** concerned with playgroups and babysitters, **topic 17** concerned with working mothers, and **topic 18** concerned with health. From the LM sub-section we focus on **Topic 7** concerned with labour market benefits and their beneficiaries, **topic 12** concerned with work contracts, **topic 16** concerned with the Dutch unemployment insurance agency, **topic 18** concerned with the general state of the economy and joblessness, and **topic 28** concerned with self-employment. We provide a more thorough topic-by-topic summary and validation of these in appendix B. The remainder of the topics are either not easily interpretable or are interpretable but not policy-relevant – this would include topics such as (automated and manual) advertising of job vacancies, advertising ECEC services, or responding to other content only with a generic emotional response (e.g. emojis denoting laughter or phrases like 'lol'). These topics are an important part of the summary of the overall information in the corpus, but are not further analyzed in this paper due to their obvious lack of policy-relevance.

From the above mentioned interpretable and potentially relevant topics we highlight a few successful examples: ECEC Topic 17 is concerned with women balancing child rearing and formal employment, which is a crucial insight for multiple policies, especially given that both sentiment and success LSS dimensions validate well for this topic: The success vs. failure dimension successfully captures how much of this content expresses dissatisfaction with this balance or with policies aimed to help this balance. ECEC topic 18 is concerned

with the health of children (and parents) at day-care and how that influences peoples' tendencies to (not) utilize day-care services. This is a very policy-relevant topic in terms of aggregating various limiting factor to take-up rate of childcare services, with the success vs. failure LSS dimension delivering a good summary of how much of this content is about policy failures and (perceived) health hazards. LM topic 28 (self-employment) is very clearly focused on various policies effecting the self-employed and the sentiment dimension provides valuable insight about how negative this content is about those policies. LM topic 12 (contracts) is a topic containing more personal commentary than other topics (generally specific to one's own employment contract), but the prevalence of both typical and a-typical working arrangements is policy-relevant.

However, there are two key issues with the policy-relevance of the identified topics. Firstly, some topics are, despite their seemingly high policy-relevance, rendered irrelevant due to factors such as strong politicisation. The prime examples of this are ECEC topic 5 and LM topic 7. Both of these topics are concerned with who are the 'beneficiaries' and 'contributors' of the Dutch welfare system, but both topics contain primarily strong anti-immigrant rhetoric. This means that despite being concerned with the redistributive effects of the two policy areas (which would be policy-relevant) the individual tweets would often bundle many different welfare policies (including but not limited to ECEC and LM) and convey that they find it unjust for immigrants to have access to these policies and for benefits to be of the size they are. Secondly, some topics remain too broad and cover multiple distinct aspects of policies in one topic, such as LM topic 28 (self-employment), LM topic 16 (unemployment insurance agency), or ECEC topic 17 (working mothers). This does not render the topics irrelevant, but limits them to a more 'agenda setting' use as they do not distinguish between individual policies (or their aspects) sufficiently. In general the summary provided here and in appendix B offers an overview of what policy-relevant information exists in our corpus, but it also points to the relative sparsity of this information among irrelevant and generic tweets and to the difficulty of summarizing this information in topics that are specific (enough) to individual policies or their aspects.

With regards to quantifying any change in these topics between the period of normalcy and of crisis we rely a) on the change in topic prevalence and b) on the change in the mean polarity score on the two LSS dimensions. The prevalence change is illustrated in Figure 7 for the four ECEC topics and in Figure 8 for the 5 LM topics. The confidence interval plotted on these figures is 95%. For the ECEC corpus sub-section we get insignificant change for topics 11 and 17. For topic 18 (health) we get a very large increase in topic prevalence and for topic 5 (ECEC benefits and beneficiaries) we see a statistically significant decrease in prevalence. For the LM corpus sub-section the only topic whose prevalence decreases is topic 28 (self-employment) with topics 7 (LM benefits and beneficiaries), 12 (work

contracts), 16 (unemployment insurance agency), and 18 (Economy and joblessness) increasing in prevalence.

With regards to change in LSS dimensions there are two significant changes in the ECEC corpus sub-section: The sentiment of topic 17 (working mothers) becomes more negative (mean change from –0.024 to –0.031) and significantly so (two-sample *t*-test *p*-value is 0.022 and Mann-Whitney *p*-value is 0.003). The success/failure dimension for topic 11 (playgroups and babysitters) becomes less associated with success (mean moving from 0.02 to 0.01) and significantly so (*p*-value for two-sample *t*-test is 0.004 and *p*-value for Mann-Whitney is 0.04). For the LM corpus sub-section the changes are the following: Topic 16 (unemployment insurance agency) experiences a very slight

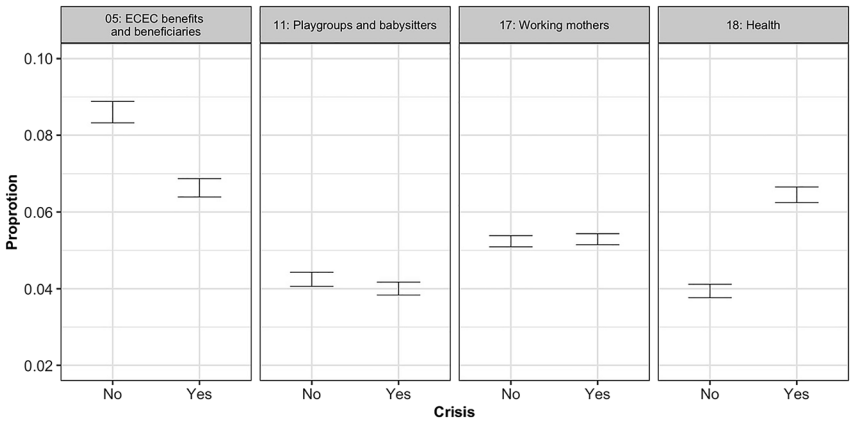


Figure 7: Change in topic-prevalence in ECEC sub-section.

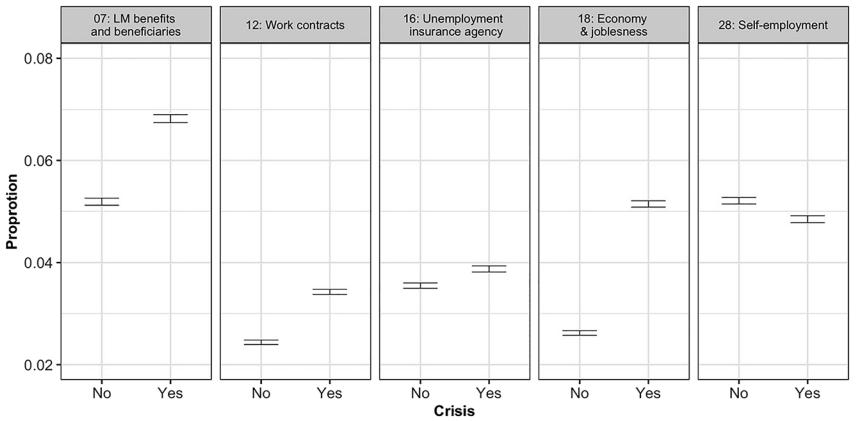


Figure 8: Change in topic prevalence between normalcy and crisis.

increase on the sentiment dimension (from −0.023 to −0.022) that is statistically significant (*t*-test *p*-value of 0.017 and Mann-Whitney *p*-value of 0.017). However, the success dimension actually becomes less concerned with failure (mean polarity changes from −0.037 to −0.033) and this change is very statistically significant (*t*-test *p*-value of 0.001 and Mann-Whitney *p*-value of 0.001). This shows that the topic becomes slightly less negative and concerned with failure, but that this change is much more prominent in its focus on failure rather than in its sentiment. Topic 12 (work contracts) actually changes substantially with regards to its sentiment (mean polarity changes from 0.001 to −0.014) and the change is statistically significant (both two-sample *t*-test and Mann-Whitney have *p*-value of 0.001). This shift is also reflected in the success dimension (a change in mean polarity from −0.004 to −0.018) at a statistically significant level (both two-sample *t*-test and Mann-Whitney have *p*-value of 0.001).

The changes in these two metrics can be read together to makes some policy-relevant observations. For example, the decreased prevalence of ECEC topic 11 (playgroups and babysitters), given that the topic is largely about existing playgroups and activities in those groups, illustrates the decline of those activities during the COVID crisis. ECEC topic 17 (working mothers) does not change in prevalence but the sentiment dimension becomes more negative during COVID crisis which illustrates that the topics itself is not 'covered' more but that its content is reflective of the increased difficulty of combining child rearing and a career with lockdown policies in place. LM topic 28 (self-employment) decreases in prevalence, potentially illustrating that lockdown policies affect the self-employed less obviously, which makes sense due to workplace shutdowns affecting mainly those with an employer. LM topic 16 (unemployment insurance agency) becomes both more prevalent and more concerned with failure, which reflects displeasure users are voicing with regards to how the agency distributes COVID-specific assistance. LM topic 12 (contracts) experiences the largest shift towards negative sentiment and towards failure observed in the data, which captures the decrease of permanent contracts being awarded (tweets celebrating obtaining such contracts are a large portion of the 'positive' tweets for that topic), which is relevant information about employment that doesn't get captured by existing indicators – people do not necessarily lose their jobs, but they seemingly experience less career advancement during the COVID-19 outbreak. That said, some of the observed change is not extremely relevant, such as the increase in prevalence of ECEC topic 18 (health) and LM topic 18 (joblessness) which is entirely expected and serve as a sanity check more so than as a finding. Some change is not as expected but still not very policy relevant, such as the decrease in prevalence of ECEC topic 5 (childcare benefits and beneficiaries) but an increase in prevalence of LM topic 7 (LM benefits and beneficiaries) despite a substantial overlap between those two topics, indicating a shift of the discussion more towards labour market issues.

Despite our ability to identify statistically significant and policy-relevant change, there are important caveats to the practical policy-relevance of these observations. For the ECEC sub-section the main caveat is a lack of data: After focusing only on tweets sufficiently representative of a topic whose sentiment is also significantly non-neutral the sample size for some topics becomes extremely small – for ECEC topic 11 we are comparing 104 tweets with 66 and for ECEC topic 17 we are comparing 48 tweets with 41. This obviously erodes the utility these observations can have in practice as the sample is extremely small, especially for leveraging the real-time nature of this data. For the LM sub-section the main caveat is the generality of topics in the sense that they contain multiple subtopics: LM topic 28 (self-employment) captures an important debate about occupational disability insurance that was very relevant in our period of 'normalcy', but also issues related to COVID relevant in period of 'crisis'. This is also the case for LM topic 16 (unemployment insurance agency) where the negative content in 'normalcy' is about discrimination and data leaks and in 'crisis' about decisions on distributing COVID-specific assistance, or ECEC topic 18 (health) that focuses on vaccinations in period of 'normalcy' and COVID testing and policies during 'crisis'. This is not an unexpected finding, but one that highlights the importance of obtaining even more disaggregated topics.

5 Limitations

There are some important limitations of our research approach and data worth noting. In terms of our data the two corpus sub-sections are very different in size and the amount of noise, but neither is able to avoid issues with both: The ECEC sub-section is smaller (by a factor of 10) and thus runs into the problem of insufficient amount of data, especially when it comes to less prevalent topics. However, this comes with the benefit of less noise (such as automated tweets or tangentially relevant commentary). The labour market sub-section is the inverse of that, as it contains much more data but at the cost of also containing a lot of noise, mainly in the form of automated tweets. In short, both sub-sections could benefit from more or better data. Some of the quality limitations of our data seem inherent to social media, and Twitter specifically, such as how opinions get formed on the platform: Moderate users tend to change their opinion over time to fit the average opinion of their friend group (Kozitsin 2020) and over time one opinion can start dominating the discourse (Xiong and Liu 2014). This adjustment of opinions is heavily influenced by 'opinion leaders' who tend to post or re-post frequently and tend to have a substantial following and be more engaged in politics (Park 2013; Park and Kaye 2017) and the strength of this influence is mediated by variables

such as trust (Xiong, Liu, and Cheng 2017) or emotion (Mansouri, Taghiyareh, and Hatami 2019). As much as users still prefer to voice their opinion rather than to change it (Xiong and Liu 2014), these dynamics are likely to over-emphasize certain opinions over others. This is not a dynamic our research design can tackle sufficiently, but there are variations of topic modelling that can potentially do so in future work (Gerrish and Blei 2010).

Outside of concerns related to data, there are also some methodological limitations of our approach. With regards to topic modelling, our approach relies on a conceptually inaccurate assumption that topics remain constant across time. This can be fixed by allowing STM models to also infer a change in the content of topics based on the crisis or normalcy variable. We do not do so for two reasons: Firstly, because models with changing topics are, in our case, less interpretable and have more diffused topics. Secondly, because we are running statistical tests of difference on the mean and median polarity, we want to keep the topic construct constant to avoid an arguably 'apples to oranges' comparison. That said, the assumption of fixed topics remains a conceptually sub-optimal assumption.

With regards to LSS the two dimensions do not validate for every topic and some topics also contain a noticeable amount of noise that can skew the LSS polarity estimate. For example, ECEC topic 11 (playgroups and babysitters) includes some comments on unrelated things like cooking that mention that the activity is happening while children are in daycare. Tweets like that can often be much more positive or negative than the more relevant policy-focused tweets, meaning that the mean for a group of tweets can be skewed by these less relevant tweets. In some cases the LSS dimension itself can be a source of bias, as some 'neutral' tokens are associated with a strong polarity in the LSS model. This happens with, for example, ECEC topic 5 where LSS models score the token 'kinderbijslag' (child benefit) itself negatively, resulting in many short tweets that mention child benefit being labeled as negative even though the sentiment of the tweet was neutral. In general we ascribe these problems to shortcomings of our data (for LSS modeling) for a few reasons: Firstly, the discourse on Twitter can be highly critical which means that even neutral tokens tend to occur in negative commentary and cause LSS to assign that token a negative polarity. Secondly, our data contains short sentences which impede the training of LSS models. We also attempted to train LSS models on entire tweets with little improvement. Thirdly, our documents (tweets) are also very short themselves and make it difficult for a trained LSS model to accurately predict a polarity as the model often only has a few words to work with (prediction is better for longer tweets in our estimation). Fourthly, the dimensions are trained on a very general corpus without restrictions to a particular topic, which results in aggregating multiple

features of any dimension (such as sentiment about weather with sentiment about politics in our case). We rely on topic belonging to minimize the effect of this (e.g. we do not expect weather and politics tweets to occupy the same topic), but that assumption is imperfect for broad topics. In general, our data is clearly sub-optimal for LSS modelling.

6 Conclusions

This paper answers the research question 'What policy-relevant information does Twitter contain?' as well as the research question 'How does this information change between a period of normalcy and a period of crisis?' For this, we analyze Twitter data for two 4-month periods – one in a period of relative normalcy and one during the first wave of the COVID-19 crisis in the Netherlands. We propose and utilize a novel method combining topic modelling with latent semantic scaling, which we design to work for real-time data streams and to easily include policymakers by giving them a comparatively large degree of control in specifying what information is 'policy-relevant'. As such, we also offer concluding remarks on the merits of this novel method.

With regards to the main research question – what policy-relevant information Twitter contains in the context of socio-economic policies – the answer is not overly optimistic: There are relevant 'topics' that include policy-relevant tweets, but these topics are relatively scarce amidst content such as job vacancy postings or responding to other tweets, which is not policy-relevant information for the purposes of this paper. Even within the relevant clusters of tweets there are substantial issues as the content can be heavily politicised, too broad in coverage, contain noise, or not validate (without a bias) on relevant LSS dimensions. Some of these problems are methodological and can likely be solved with more sophisticated topic modelling or data cleaning methods, but some of these problems (like the heavy politicisation of some of the relevant content) are not a methodological flaw and are simply a feature of our data. This results in most of the identified relevant topics likely only having utility for agenda setting and not for monitoring performance of specific policies. Furthermore, some of these tweet clusters are policy-relevant according to our definition, but their extremely small size makes them of no practical utility for policymaking.

The issues we run into with our data and method for extracting insight are seemingly related to features of Twitter data that necessitate us to make trade-offs that we would ideally avoid entirely. The central issue is that it is impossible to perfectly distinguish between signal and noise: Despite removing

re-tweets, accounts posting with high frequency, and duplicate tweets, near-duplicate tweets still exist due to various sharing features and differences in '@' mentions. If we pre-process data more aggressively, we start losing a noticeable amount of 'signal'. Our approach of conceding an amount of 'noise' in the corpus and letting it cluster into topics that can be excluded works well (in our estimation), but certainly not perfectly. Similar trade-offs exist for our treatment of '@' mentions (we do not remove them and in doing so minimize information loss but it has implications for our topic models), re-tweets (we exclude them to prioritize original and personal content but lose the associated information), quoted tweets (we join the tweet text with the text of the tweet it quotes which introduces noise but maintains the context of those tweets), and others. For any of these processing decisions our choice is informed, but at the same time far from perfect, which aggregates into meaningful problems for our analysis. We consider this as a part of our findings relevant to understanding the limitations of using twitter data in policymaking.

With regards to the second research question the answer is, despite being plagued by the same issues mentioned above, much more straightforward. We show that the prevalence (how much is something being talked about) as well as what is being said (in terms of LSS dimensions) changes for some topics and remains consistent for others and that this (lack of) change is statistically significant at a 95% confidence level. Some of these changes are entirely expected, such as the increase in prevalence of topics concerned with the impact of corona on health of children or on the economy. Some findings are a bit more interesting and can carry a great deal of policy relevance, such as the changes to ECEC topic 11 (playgroups and babysitters), ECEC topic 17 (working mothers), LM topic 28 (self-employment), or LM topic 12 (contracts) that we highlight in Sub-section 4.3. Some of these insights, such as the decrease in tweets celebrating the obtaining of permanent (and generally better) contracts, capture information that is both policy-relevant and not easily obtainable from other data sources. That said, these changes are only as informative as the topics themselves, which constitutes an important limitation given the general difficulty of identifying relevant topics and their (often insufficient) size.

Even though our findings point to serious limitations of utilizing Twitter data for policymaking, our conclusion should not be viewed as contradictory to other research that finds Twitter data useful in the context of labour market indicators (e.g. Proserpio, Counts, and Jain (2016), or in the context of the COVID-19 crisis (e.g. Gilardi et al. (Forthcoming)). Our conclusion is complementary to those findings due to differences in approach and sampling strategies. Our approach is distinct in that we a-priori select two rather broad policy domains and collect all relevant tweets without controlling for users or a strict

policy focus, unlike research efforts collecting data for specific users and gathering a more 'complete' profile for those users (Gilardi et al. Forthcoming; Proserpio, Counts, and Jain 2016). There are of course also important methodological differences but taken together with the existing literature our findings point to a potential limit to the utility of Twitter data.

Despite this cautionary conclusion, the proposed method performed relatively well, especially given the relative brevity of tweets, the fact that some dimensions are not applicable to all topics (some topics are simply not concerned with, for example, failure or success), or the skew of topics themselves (some topics seemingly do not contain a lot of positive sentiment or talk of success). As a general method it has a number of comparative advantages in the level of detail it provides, and the degree of control researchers or policymakers have when it comes to defining dimensions of interest. Technically the method is also language independent and able to utilize streaming data in a real-time fashion with periodical re-training of the underlying models. Needless to say, more future work is needed to validate our approach as it is currently not validated to ground-truth data about public perception such as general opinion surveys or surveys of users whose tweets are included in our corpus. Beyond validation on more fitting data substantial work also remains with regards to seeing how policymakers would utilize such a system and how well it would perform compared to existing decision support systems.

Data availability statement: Data cannot be shared due to terms and conditions for using Twitter's API. Given the recent changes to Twitter's API the keywords in Appendix A can aid (partial) replication.
Method replication statement: Code necessary to replicate the full method can be found at https://github.com/SimonVydra/STM-LSS.

Appendix A Keywords

All included keywords are utilized in gathering data from Twitter's API. This general corpus is utilized to train LSS models.

The ECEC sub-section of the corpus includes only tweets that contain at least one keyword from the 'ECEC' section below in the text of the tweet.

The LM sub-section of the corpus includes only tweets that contain at least one keyword from the 'Labour Market' section, with the notable exception of all keywords that are crossed out. The crossed out keywords are excluded for making corpus sub-sections but are included for LSS training.

ECEC:

'kinderopvang' 'kinder opvang'	General childcare term
'kinderdagverblijf' 'kdv'	Refers to daycare centers Cover children up to four years old and 10 h a day during working hours. Apparently limited capacity
'gastouder' 'gastouders' 'gastouderopvang' 'gastouder opvang' 'gastouderbureau'	Available for toddlers up to pre-school by parents caring for up to six children in locations approved by the national childcare register. Often administered by agencies
'peuterspeelzalen' 'peuterspeelzaal' 'peuterspeelplaats'	Preschools, these are usually part of a primary school and are a preparatory program for children between two and four. These do not cover the whole week or full days.
'peutergroep' 'peutergroepen'	A more informal playgroup setting for young children
'buitenschoolseopvang' 'buitenschoolse opvang' 'naschoolseopvang' 'naschoolse opvang' 'naschoolse' 'BSO' 'voorschoolse opvang' 'voorschoolse' 'voorschoolseopvang'	Afterschool and outdoor school care. However, this is connected to primary schools and thus generally available to children from 4 years of age
'oppas' 'oppassers' 'babysitter' 'babysitters' 'nanny' 'nannies'	Babysitter options
'kinderopvangtoeslag'	Childcare subsidy in the Netherlands
'kindgebonden budget'	Automatic child benefit if your child is under 18 and your income is not high
'kinderbijslag'	Covers part of the cost of raising children and depends on their number and residence

Labour market:

Legislation and programs:

'Participatiewet' 'Participatie wet' 'Gesubsidieerde arbeid'	Overarching legislation in place to support people who can work but need some sort of assistance in order to work.
~~'opleiding' 'scholing' 'heropleiding' 'omscholing' 'training' 'retraining' 're-training' 'studie' 'studeren'~~	~~Training and re-training.~~
~~'praktijktraining' 'werkervaringsplek' 'stage' 'stage lopen' 'werkervaringsplek' 'werkervaring plek' 'studeer en werkplek' 'studeer- en werkplek' 'traineeship'~~	~~On the job training apprenticeship~~
'Werkbedrijf' 'werk.nl' 'werkplein' 'werkpleinen' 'arbeidsadviseur' 'uwv' 'arbeidsbemiddelaar' 'arbeidsbemiddeling' 'loopbaan coach' 'werk coach'	Employment services
'WW-uitkering' 'uitkering' 'bijstand' 'bijstandsuitkering' ~~'meewerkaftrek'~~	Unemployment (benefits) ~~Subsidy for when your partner works in your business without pay~~

Type of employment:

'full-time werk' 'full time werk' 'fulltime werk' 'full-time baan' 'full time baan' 'fulltime baan' 'voltijd baan' 'voltijd werk' 'voltijdwerk' '1 fte' '1 wtf'	Full time work
'deeltijd werk' 'part-time werk' 'part time werk' 'deeltijd baan' 'part-time baan' 'part time baan'	Part time work
'vast contract' 'vaste baan' 'vaste aanstelling'	Permanent contract
'tijdelijk contract' 'tijdelijke baan' 'tijdelijke aanstelling'	Temporary contract
'uitzendcontract'	Contract with recruitment agency
'nul uren contract' '0 uren contract'	Zero hour contract
'zelfstandige zonder personeel' 'zzp' 'zzp'ers' 'zzp'er' 'zzper' 'zzpers' 'DBA modelovereenkomst'	freelancers
'schijnzelfstandigheid'	Sham independence
'loondienst' 'in loondienst'	Salaried employment
'eigen baas' 'eigen baas zijn'	Self-employment

Generic employment-related phrases:

'werkloosheid' 'werkeloosheid' 'werkloos' 'zonder baan' 'jobless' 'in between jobs' 'between jobs' 'in between two jobs' 'between two jobs'	Unemployment
'onderbezetting' 'onderbezet'	Underemployment
'zoek naar werk' 'kijken voor werk' 'een baan zoeken' 'zoeken naar een baan' 'banen zoeken'	Job search
'passend werk' 'passende arbeid' 'passende baan' 'passende job'	Correct or fitting job
'goed werk' 'slecht werk' 'beter werk' 'betere kansen op werk' 'beter arbeidscontract' 'goed arbeidscontract' 'slecht arbeidscontract'	A good job
'vacature' 'vacatures' 'openstaande baan'	Job vacancies
'vaardigheidseisen' 'ervaringseisen' 'werkervaring' 'werkervaringseisen' 'competenties'	Skill/experience requirements

Appendix B Topic Summaries

In this appendix we offer a matter-of-fact summary of selected topic models for both corpus sub-sections. Section B.1 focuses on early childhood education and care corpus sub-section and Section B.2 focuses on the labour market sub-section. We provide summaries of the entire model (in English and Dutch) and a topic-by-topic description of all potentially policy-relevant and interpretable topics

including insight from manual inspection of the topics and LSS dimension scores for individual tweets.

B.1 ECEC Corpus

For the ECEC sub-section of the corpus, the most interpretable model is the 20-topic model, delivering similar topics to the 25 and 30-topic models but with more interpretability and more 'focused' topics. All topics of this model (described by top 10 tokens) are summarized in Figures B1 (English translation) and B2 (Dutch original). For Figures B1–B4 we sort top tokens by probability to appear for a topic, which is inferred directly from the topic-word distribution. There are alternative metrics to sort top words by, but in this case we find this to represent the topics most accurately.

The interpretable topics in this case are the following: **Topic 5** is defined by tokens like 'child benefit', 'benefit' (toeslag), 'payment', 'to pay', 'money', 'budget', or 'country' and generally focuses on who receives what benefits and who is paying for them. This makes the topic not very exclusive to childcare and also concerned with political issues like the support migrants receive. The sentiment and success LSS dimensions are both negative, but mainly reveal a bias: The models score the token 'kinderbijslag' (child benefit) itself negatively, resulting in many short tweets that mention it as being highly negative and related to failure. This results in neither dimension validating.

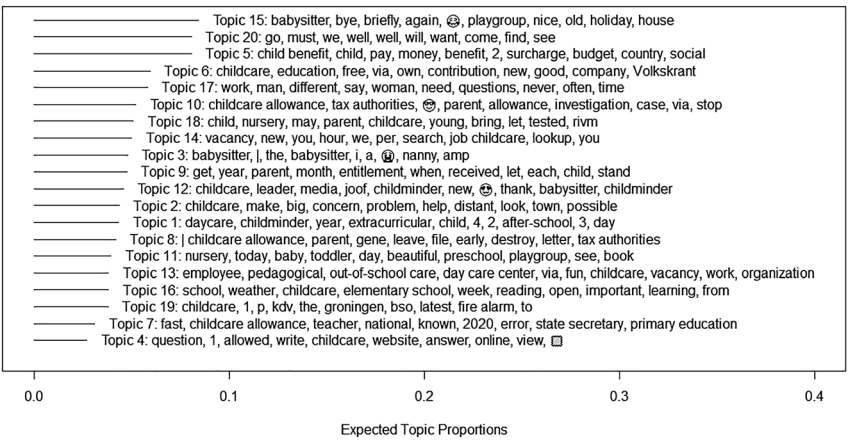


Figure B1: Topics in a 20-topic model of ECEC sub-section (English translation).

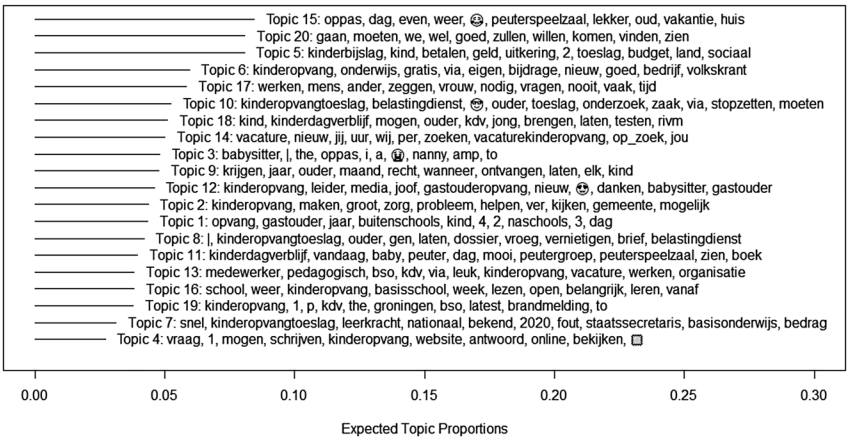


Figure B2: Topics in a 20-topic model of ECEC sub-section (Dutch original).

Topic 17 is defined by tokens like 'to work', 'people', 'women', 'time', 'necessary', 'often', or 'never' and generally focuses on issues related to women's role in the labour market and (unpaid) domestic work. It includes commentary on the choice between paid or unpaid labour and family life both from the perspective of women stating and justifying their choices and from a perspective of more general commentary. For the sentiment dimension the mean polarity is negative (–0.027) and statistically significant. The dimension validates, with negative polarity corresponding to people arguing and insulting one another and positive polarity corresponding to people looking for babysitters or commenting on their choices with regards to child rearing and employment in neutral or positive terms. The success/failure dimension also validates and is negative (mean –0.027), with the failure polarity remaining 'negative' but focused more on failure of policy or individual providers rather than insults.

Topic 18 is defined by tokens like 'child', 'parent', 'to bring', 'childcare' (both as full word and 'kdv'), 'to test', or 'rivm' (National Institute for Health and Environment). In top 20 tokens words like 'sick' or 'corona' also appear. The topic generally focuses on health in childcare, with a detectable focus on COVID-19 and vaccinations. Only the success/failure dimension validates here – mean polarity on this dimension is negative (–0.024) and significant, with the negative polarity being about necessary policy adjustments and various failures of policies, including some general negative commentary.

Topic 11 is defined by tokens like 'childcare', 'playgroup' (multiple tokens), 'today', 'day', 'nice', or 'to see' and it is a somewhat general topic about childcare and playgroups that contains personal commentary on playgroups or preschools,

pointing to newspaper articles about preschools, or preschools advertising themselves. The sentiment dimension validates but shows the generality of the topic as some positive tweets are only tangentially relevant. The mean polarity of the sentiment dimension is slightly positive (0.012) and statistically significant. The success/failure dimension also validates well with the failure polarity focusing on failure, insufficiency, and general hazard and the success polarity focusing on success. The mean polarity is slightly towards success (0.016).

B.2 LM Corpus Sub-section

For the Labour Market sub-section, the most interpretable model is the 30-topic model. Similar to the ECEC sub-section, this model contains generally the same interpretable topics, but seemingly includes less noise for those topics than other candidate models. All topics of this model (described by top 10 tokens) are summarized in Figures B3 (English translation) and B4 (Dutch original).

The interpretable topics in this case are the following: **Topic 7** includes tokens like 'payment', 'assistance', 'people', 'receive', 'must', 'money', or 'work' and describes who is receiving benefits, what type of benefits, how much they total to,

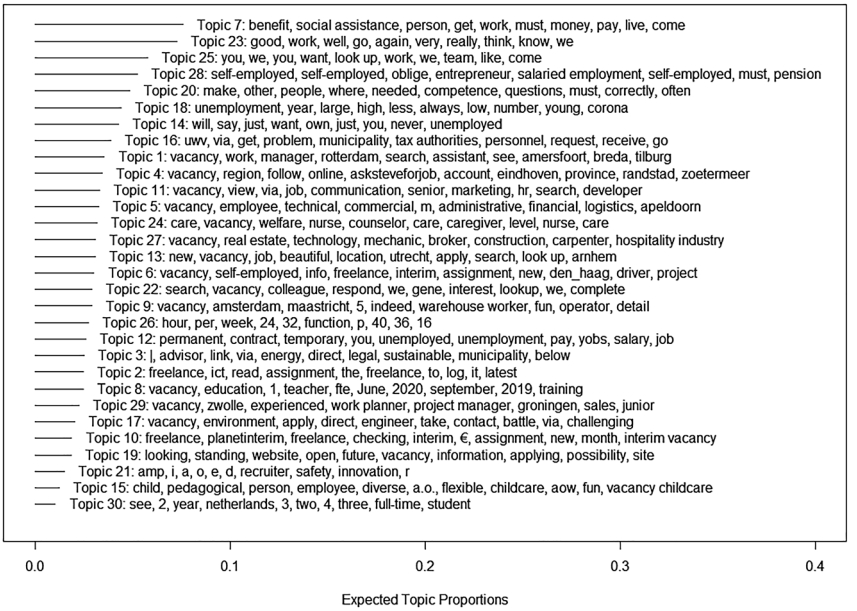


Figure B3: Topics in a 30-topic model of LM sub-section (English translation).

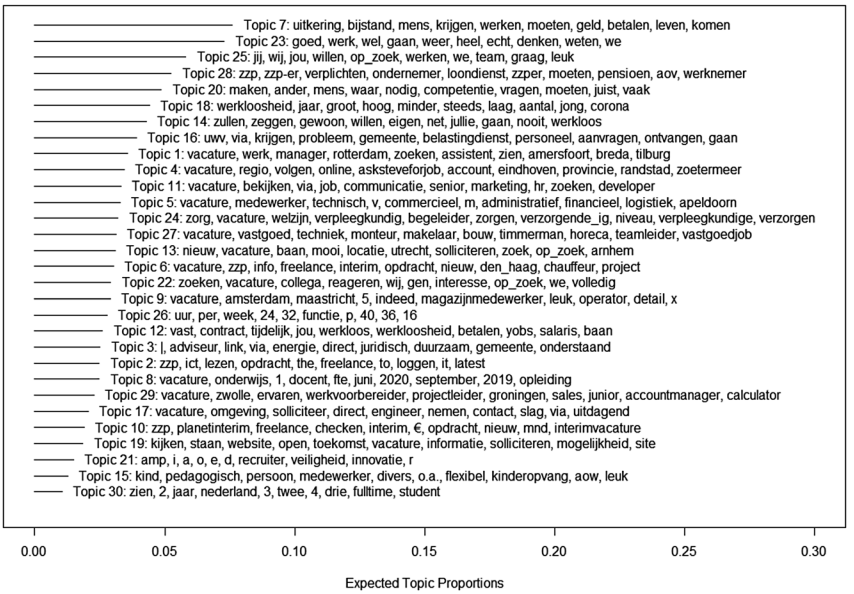


Figure B4: Topics in a 30-topic model of LM sub-section (Dutch original).

and whether it is deserved. This topic is heavily concerned with immigration, and there is a strong overlap between this topic and topic 5 from the ECEC sub-section. In terms of sentiment the dimension validates, but with a noticeable bias: The positive extreme is generally slightly positive or neutral and the negative extreme shows noticeable bias due to the word 'uitkering' (payment) being labeled as negative on this dimension. The mean sentiment polarity is slightly negative (–0.03) and statistically significant. The success/failure dimension doesn't validate.

Topic 28 includes tokens like 'self-employment', 'self-employed (noun)', 'entrepreneur', 'oblige', 'pension', 'occupational disability insurance (aov)' and describes various aspects of policy prescriptions for the self-employed. These include pension, the legal distinction between self-employed and entrepreneurs, but mainly whether the self-employed should have to contribute to occupational disability insurance. The sentiment dimension validates well, with the positive extreme praising policy changes or stating that policy is finally negotiated and passed, and the negative dimension mentioning that something is being critiqued or perceived negatively. The mean sentiment polarity is slightly negative (–0.02) and significant. The success dimension validates with mean polarity towards failure (–0.03) and significant.

Topic 18 includes tokens like 'joblessness', 'high', 'big', 'year', 'less', 'corona', 'economy' and is clearly a topic providing commentary about joblessness and the economic (and overall) impact of corona. However, it doesn't differentiate between the situation in the Netherlands and elsewhere in the world, including commentary on the US or the Eurozone. Sentiment dimension doesn't validate due to un-interpretable positive polarity, and success dimension doesn't validate due to not interpretable success polarity.

Topic 16 includes tokens like 'unemployment insurance agency', 'to receive', 'problem', 'municipality', 'tax authority', or 'via' (often used to link to a news story). It is mainly concerned with issues relevant to the UVW (unemployment insurance agency) like data leaks, miscalculations, or the misuse of corona-specific assistance. The sentiment dimension validates fine with negative polarity clearly corresponding to negative and critical comments, and positive polarity containing neutral or slightly positive commentary. The mean sentiment polarity is slightly negative (–0.02) and significant. The success dimension validates very well, with the failure polarity associated with tweets that are about failures and shortcomings of policies or the UVW, and the success polarity being much more neutral but reliably excluding comments about blatant failure. The polarity here is towards failure (–0.035) and is significant.

Topic 12 includes tokens like 'fixed', 'temporary', 'contract', 'joblessness', 'jobless', 'to pay', 'salary' and is rather general but maintains a focus on type of contracts. The issues this covers range from technical errors in the administration's systems, to family postponement due to non-fixed contracts. Sentiment validates despite the negative polarity being largely dispassionate and the positive polarity including some noise. The sentiment polarity is very slightly negative (–0.006) but still significant due to large sample. The success dimension is similar with success polarity corresponding to people celebrating new employment contracts and the failure polarity corresponding to negative comments concerned primarily with unemployment. The success dimension leaning towards failure slightly (–0.01) and is significant.

References

Acar, A., and Y. Muraki. 2011. "Twitter for Crisis Communication: Lessons Learned from Japan's Tsunami Disaster." *International Journal of Web Based Communities* 7: 392–402. Inderscience Publishers.

Ahmed, W., P. A. Bath, L. Sbaffi, and G. Demartini. 2019. "Novel Insights into Views towards H1N1 during the 2009 Pandemic: a Thematic Analysis of Twitter Data." *Health Information and Libraries Journal* 36: 60–72. Blackwell Publishing Ltd.

Andriotis, P., and A. Takasu. 2019. "Emotional Bots: Content-Based Spammer Detection on Social Media." In *10th IEEE International Workshop on Information Forensics and Security*, 1–8. WIFS 2018, IEEE. https://doi.org/10.1109/WIFS.2018.8630760.

Antenucci, D., M. Cafarellab, M. Levensteinc, C. Red, and M. Shapiro. 2014. *Using Social Media to Measure Labor Market Flows*. NBER Working Paper No. 20010.

di Bella, E., L. Leporatti, and F. Maggino. 2018. "Big Data and Social Indicators: Actual Trends and New Perspectives." *Social Indicators Research* 135: 869–78. Springer Netherlands.

Baker, S., and A. Fradkin. 2011. *What Drives Job Search? Evidence from Google Search Data*. Stanford Institute for Economic and Policy Research Working Paper No. 10–020.

Biorci, G., A. Emina, M. Puliga, L. Sella, and G. Vivaldo. 2017. "Tweet-Tales: Moods of Socio-Economic Crisis?" In *Data Science and Social Research. Studies in Classification, Data Analysis, and Knowledge Organization*, edited by N. Lauro, E. Amaturo, M. Grassia, B. Aragona, and M. Marino. Cham: Springer. https://doi.org/10.1007/978-3-319-55477-8_19.

Blei, D. M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 77–84, https://doi.org/10.1145/2133806.2133826.

Blei, D. M., and J. D. Lafferty. 2006. "Dynamic Topic Models." In *Proceedings of the 23rd International Conference on Machine Learning*, 113–20. New York, NY, USA: ACM Press. https://doi.org/10.1145/1143844.1143859.

Blei, D. M., and J. D. Lafferty. 2007. "A Correlated Topic Model of Science." *Annals of Applied Statistics* 1: 17–35.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Carta, F., and L. Rizzica. 2016. "Female Employment and Pre-kindergarten: On the Unintended Effects of an Italian Reform." *Dondena Working Papers*, 91.

Cavallo, A., E. Cavallo, and R. Rigobon. 2014. "Prices and Supply Disruptions during Natural Disasters." *Review of Income and Wealth* 60 (2): 449–71.

CBS. 2020. "Monthly Labour Participation and Unemployment." Available at https://www.cbs.nl/en-gb/figures/detail/80590eng.

Chen, G. M. 2011. "Tweet This: A Uses and Gratifications Perspective on How Active Twitter Use Gratifies a Need to Connect with Others." *Computers in Human Behavior* 27: 755–62. Elsevier Ltd.

Chetty, R., J. N. Friedman, N. Hendren, M. Stepner, and OI Team. 2020. *The Economic Impacts of COVID-19: Evidence from a New Public Database Built from Private Sector Data*. NBER Discussion paper, no.2743.

Chew, C., and G. Eysenbach. 2010. "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak." *PloS One* 5: e14118. (M. Sampson, ed.), Public Library of Science.

Dwivedi, Y. K., N. P. Rana, M. Tajvidi, B. Lal, G. P. Sahu, and A. Gupta. 2017. "Exploring the Role of Social Media in E-Government: An Analysis of Emerging Literature." In *ACM International Conference Proceeding Series*, 97–106. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3047273.3047374.

Emmanuel, I., and C. Stanier. 2016. "Defining Big Data." In *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies – BDAW '16*, 1–6. New York, NY, USA: ACM Press. https://doi.org/10.1145/3010089.3010090.

European Commission. 2020. "LMP Expenditure by Type of Action." Available at https://webgate.ec.europa.eu/empl/redisstat/databrowser/view/LMP_EXPSUMM/default/table?lang=en&category=lmp_expend.

Eurostat. 2020. "Children in Formal Childcare or Education by Age Group and Duration." Available at https://ec.europa.eu/eurostat/databrowser/view/ILC_CAINDFORMAL__custom_129524/default/table?lang=en.

Gallego, A., and P. Marx. 2017. "Multi-dimensional Preferences for Labour Market Reforms: a Conjoint Experiment." *Journal of European Public Policy* 24: 1027–47. Routledge.

Gerrish, S. M., and D. M. Blei. 2010. "A Language-Based Approach to Measuring Scholarly Impact." In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 375–82. Omnipress.

Gilardi, F., T. Gessler, M. Kubli, and S. Müller. Forthcoming. *Social Media and Policy Responses to the COVID-19 Pandemic in Switzerland*. Swiss Political Science Review.

Grammatikopoulos, V., A. Gregoriadis, N. Tsigilis, and E. Zachopoulou. 2014. "Parental Conceptions of Quality in Greek Early Childhood Education." *European Early Childhood Education Research Journal* 22: 134–48.

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21: 1–31.

Grubmüller, V., K. Götsch, and B. Krieger. 2013. "Social Media Analytics for Future Oriented Policy Making." *European Journal of Futures Research* 1, https://doi.org/10.1007/s40309-013-0020-7.

Gualda, E., and C. Rebollo. 2016. "The Refugee Crisis on Twitter: A Diversity of Discourses at A European Crossroads." *Journal of Spatial and Organizational Dynamics* 4: 199–212.

Heverin, T., and L. Zach. 2010. "Microblogging for Crisis Communication: Examination of Twitter Use in Response to a 2009 Violent Crisis in the Seattle-Tacoma, Washington Area,." In *Proceedings of the 7th International ISCRAM Conference*. Seattle.

Höchtl, J., P. Parycek, and R. Schöllhammer. 2016. "Big Data in the Policy Cycle: Policy Decision Making in the Digital Era." *Journal of Organizational Computing & Electronic Commerce* 26: 147–69. Taylor & Francis.

ILO. 2020. *ILO Monitor: COVID-19 and the World of Work*, 6th ed. ILO.

Internet World Stats. 2017. "Europe Internet Usage Stats Facebook Subscribers and Population Statistics." Available at http://www.internetworldstats.com/stats4.htm.

Inuwa-Dutse, I., M. Liptrott, and I. Korkontzelos. 2018. "Detection of Spam-Posting Accounts on Twitter." *Neurocomputing* 315: 496–511. Elsevier.

Janssen, M., and N. Helbig. 2018. "Innovating and Changing the Policy-Cycle: Policy-Makers Be Prepared!" *Government Information Quarterly* 35: S99–S105. Elsevier Ltd.

Java, A., X. Song, T. Finin, and B. Tseng. 2007. "Why We Twitter: Understanding Microblogging Usage and Communities." In *Procedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007*. Association for Computing Machinery.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science, American Association for the Advancement of Science* 353: 790–4.

Johnson, P. R., and S.-U. Yang. 2009. "Uses and Gratifications of Twitter: An Examination of User Motives and Satisfaction of Twitter Use." In *Paper Presented at the Communication Technology Division of the Annual Convention of the Association for Education in Journalism and Mass Communication*. Available at https://www.researchgate.net/publication/

228959109_Uses_and_gratifications_of_Twitter_An_examination_of_user_motives_and_satisfaction_of_Twitter_use.

Kantepe, M., and M. C. Ganiz. 2017. "Preprocessing Framework for Twitter Bot Detection." In *2017 International Conference on Computer Science and Engineering (UBMK)*, 630–4. London: IEEE.

Kawabata, M. 2012. "Access to Childcare and the Employment of Women with Preschool-Aged Children in Tokyo." *CSIS Discussion Paper*, 114.

Kozitsin, I. V. 2020. "Formal Models of Opinion Formation and Their Application to Real Data: Evidence from Online Social Networks." *Journal of Mathematical Sociology*, https://doi.org/10.1080/0022250X.2020.1835894.

Lee, K., B. Eoff, and J. Caverlee. 2011. "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter." In *ICWSM*. PKP.

Mansouri, A., F. Taghiyareh, and J. Hatami. 2019. "Improving Opinion Formation Models on Social Media through Emotions." In *5th International Conference on Web Research (ICWR)*, 6–11. IEEE.

Matsaganis, M., E. Ozdemir, and T. Ward. 2014. *The Coverage Rate of Social Benefits*. European Commission.

McDaniel, B. T., S. M. Coyne, and E. K. Holmes. 2012. "New Mothers and Media Use: Associations between Blogging, Social Networking, and Maternal Well-Being." *Maternal and Child Health Journal* 16: 1509–17. Springer US.

McNeill, A., P. R. Harris, and P. Briggs. 2016. "Twitter Influence on UK Vaccination and Antiviral Uptake during the 2009 H1N1 Pandemic." *Frontiers in Public Health, Frontiers Media S.A.* 4: 22.

Mergel, I., and S. I. Bretschneider. 2013. "A Three-Stage Adoption Process for Social Media Use in Government." *Public Administration Review* 73: 390–400. John Wiley & Sons, Ltd.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. Mccallum. 2011. "Optimizing Semantic Coherence in Topic Models." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh: Association for Computational Linguistics.

Öztürk, N., and S. Ayvaz. 2018. "Sentiment Analysis on Twitter: A Text Mining Approach to the Syrian Refugee Crisis." *Telematics and Informatics* 35: 136–47. Elsevier Ltd.

Panagiotopoulos, P., F. Bowen, and P. Brooker. 2017. "The Value of Social Media Data: Integrating Crowd Capabilities in Evidence-Based Policy." *Government Information Quarterly, JAI* 34: 601–12.

Park, C. S. 2013. "Does Twitter Motivate Involvement in Politics? Tweeting, Opinion Leadership, and Political Engagement." *Computers in Human Behavior* 29: 1641–8. Pergamon.

Park, C. S., and B. K. Kaye. 2017. "The Tweet Goes on: Interconnection of Twitter Opinion Leadership, Network Size, and Civic Engagement." *Computers in Human Behavior* 69: 174–80. Elsevier Ltd.

Proserpio, D., S. Counts, and A. Jain. 2016. "The Psychology of Job Loss: Using Social Media Data to Characterize and Predict Unemployment." In *Proceedings of the 8th ACM Conference on Web Science*, 223–32. New York, NY, USA: ACM Press. https://doi.org/10.1145/2908131.2913008.

Prpić, J., A. Taeihagh, and J. Melton. 2015. "The Fundamentals of Policy Crowdsourcing." *Policy & Internet* 7: 340–61. Wiley-Blackwell Publishing Ltd.

Roberts, M. E., B. M. Stewart, and E. M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111: 988–1003. American Statistical Association.

Roberts, M. E., B. M. Stewart, and D. Tingley. 2019. "Journal of Statistical Software Stm: R Package for Structural Topic Models." *Journal of Statistical Software* 91, https://doi.org/10.18637/jss.v000.i00.

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58: 1064–82. Blackwell Publishing Ltd.

Salganik, M. J. 2018. *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Signorini, A., A. M. Segre, and P. M. Polgreen. 2011. "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. During the Influenza A H1N1 Pandemic." *PloS One* 6: e19467. (A. P. Galvani, ed.), Public Library of Science.

Singh, P., Y. K. Dwivedi, Karanjeet, S. Kahlon, Ravinder, S. Sawhney, A. A. Alalwan, and N. P. Rana. 2020. "Smart Monitoring and Controlling of Government Policies Using Social Media and Cloud Computing." *Information Systems Frontiers* 22: 315–37.

Statista. 2018. "Netherlands: Twitter Users, by Age Group 2017-2018." Available at https://www.statista.com/statistics/828876/twitter-penetration-rate-in-the-netherlands-by-age-group/.

Statista. 2019. "Netherlands: Number of Twitter Users 2013-2019." Available at https://www.statista.com/statistics/880865/number-of-twitter-users-in-the-netherlands/.

Szomszor, M., P. Kostkova, and E. De Quincey. 2011. "#Swineflu: Twitter Predicts Swine Flu Outbreak in 2009." In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, 18–26. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-23635-8_3.

Taddy, M. A. 2012. "On Estimation and Selection for Topic Models." In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, 1184–93. Association for Computing Machinery.

Terpstra, T., A. de Vries, R. Stronkman, and G. L. Paradies. 2012. "Towards a Real time Twitter Analysis during Crises for Operational Crisis Management." In *Proceedings of the 9th International ISCRAM Conference*. Vancouver: Simon Fraser University.

Turrell, A., B. J. Speigner, J. Djumalieva, D. Copple, and J. Thurgood. 2019. "Transforming Naturally Occurring Text Data into Economic Statistics: The Case of Online Job Vacancy Postings." In *Big Data for 21st Century Economic Statistics*, edited by K. G. Abraham, R. S. Jarmin, B. Moyer, and M. D. Shapiro. Chicago: University of Chicago Press.

United Nations. 2011. "Unemployment through the Lens of Social Media." Available at http://www.unglobalpulse.org/projects/can-social-media-mining-add-depth-unemployment-statistics.

Van Den Bosch, A., B. Busser, S. Canisius, and W. Daelemans. 2007. "An Efficient Memory-Based Morphosyntactic Tagger and Parser for Dutch." In *Computational Linguistics in the Netherlands*, edited by F. V. Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste, 191-206. Utrecht : LOT.

Vydra, S., and B. Klievink. 2019. "Techno-optimism and Policy-Pessimism in the Public Sector Big Data Debate." *Government Information Quarterly* 36, https://doi.org/10.1016/j.giq.2019.05.010.

Wang, Y., E. Agichtein, and M. Benzi. 2012. "TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media." In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 123–31. New York, NY, USA: ACM Press. https://doi.org/10.1145/2339530.2339552.

Ward, J. S., and A. Barker. 2013. "Undefined by Data: A Survey of Big Data Definitions." arXiv: 1309.5821.

Watanabe, K. 2017a. "The Spread of the Kremlin's Narratives by a Western News Agency during the Ukraine Crisis." *Journal of International Communication* 23: 138–58. Taylor and Francis.

Watanabe, K. 2017b. "Measuring News Bias: Russia's Official News Agency ITAR-TASS' Coverage of the Ukraine Crisis." *European Journal of Communication* 32: 224–41. SAGE Publications Ltd.

Watanabe, K. 2020. "Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages." *Communication Methods and Measures*, https://doi.org/10.1080/19312458.2020.1832976.

Xiong, F., and Y. Liu. 2014. "Opinion Formation on Social Media: an Empirical Approach." *Chaos* 24: 013130. (Woodbury, N.Y.), American Institute of Physics AIP.

Xiong, F., Y. Liu, and J. Cheng. 2017. "Modeling and Predicting Opinion Formation with Trust Propagation in Online Social Networks." *Communications in Nonlinear Science and Numerical Simulation* 44: 513–24. Elsevier B.V.

Ylijoki, O., and J. Porras. 2016. "Perspectives to Definition of Big Data: A Mapping Study and Discussion." *Journal of Information Management* 4 (1): 69–91.