Utrecht University

Master's thesis

Master Applied Data Science

# Evaluating Topic Tracking and Trend Identification Methods for Finding Trending Topics in Open-Ended Healthcare Policy Questionnaire Responses

Author: Jop Gelinck

Student number: 0115754

First supervisor: Dr. Florian Kunneman

Second supervisor: Prof. dr. Antal van den Bosch

July 7th, 2024

Universiteit Utrecht

# Table of Contents

# 1 Abstract

This thesis evaluates the effectiveness of various topic tracking and trend identification methods in detecting trending topics in open-ended healthcare policy questionnaire responses. Identifying these trending topics is crucial for providing decision-makers with insights into public concerns and important policy areas.

Two primary model types are used in this study. First, topic tracking models, including Length Weighted Topic Chains (LWTC) and Single Pass Clustering (SPC), are utilized to identify topics and generate a timeseries of their occurrences. Second, trend identification models leverage these timeseries to detect trending topics through methods such as Most Occurring Topics Selection, Moving Average (MA), and AutoRegressive Integrated Moving Average (ARIMA), or by analyzing word frequency using Bursty Term Extraction (BTE).

The evaluation of topic tracking methods is conducted via the word intrusion task, while trend identification models are assessed using precision and F1-scores across two labeled timelines. The findings indicate that SPC beats LWTC in topic modeling but lacks generalizability, whereas LWTC demonstrates consistent performance. For trend identification, the Moving Average method emerges as the most effective, achieving the highest precision and F1-score combination, followed by ARIMA, Most Occurring Topics Selection, and BTE.

A notable result is that the combination of LWTC with the Moving Average method yields the best overall performance for identifying trending topics in open-ended healthcare policy questionnaire responses. This combination achieves both a high precision and F1-scores, making it the most robust approach in this context. However, the thesis also reveals a disconnect between topic modeling quality and trend detection effectiveness, suggesting that higher interpretability does not necessarily translate to superior trend identification.

This thesis highlights several limitations, including the subjectivity in topic labeling and comparison with labeled timelines, and the lack of expert involvement in evaluation tasks. Future studies should address these by incorporating domain experts and developing expert-based evaluation metrics to enhance the practical utility and accuracy of these methods. Additionally, improvements in topic modeling performance for LWTC through BERTopic or Biterm Topic Model (BTM) could further refine trending topic identification.

In conclusion, this thesis contributes valuable insights into the application of topic tracking and trend identification models in healthcare policy analysis, offering a promising approach for extracting valuable information from open-ended questionnaire responses.

# 2 Introduction

## 2.1 Motivation

Periodic questionnaires about health care policies can provide various insights for a national health organization. They can give insight into the perception and awareness of the public on subjects such as accessibility and quality of health care, the effectiveness and priorities of policies and how well these align with public sentiment. Additionally, they can generate meaningful insight into certain issues that may be on the rise, such as an outbreak or rise in cases of a disease. By analyzing the trends of the topics mentioned in these questionnaires, adaptations to the policies and the motivations for them can be made to better service the general public. Additionally, the information provided to the general public can be increased or improved to better inform the public about certain subjects.

However, to analyze these topical trends, they must be extracted from the questionnaires. These questionnaires consist of a mix of closed and open questions, and whereas the closed questions can easily be analyzed, it is harder to easily extract meaningful information from open question text fields. This is especially the case in large quantities of filled out questionnaires. Analyzing these open question fields of the questionnaires by hand would take a very long time and a lot of manpower. Machine Learning (ML) and Natural Language Processing (NLP) approaches can drastically cut the cost and time required for analyzing this data (Guetterman et al., 2018). An example of why reducing the time and man hours needed for the analysis is so important is the COVID-19 pandemic. During this time, national healthcare agencies had too few people and needed to make decisions fast. In such a case, ML and NLP approaches would allow for timely analysis of topical trends in these questionnaires.

And while all kinds of topical trends contain valuable information, such as a topic getting more interest, less interest or even remaining stagnant over time, identifying topics that are new or gather a lot more interest should be the number one priority. This is because quickly identifying trending or new topics is crucial in health care policy since it enables timely responses to emerging issues, which can significantly impact public health and safety. When a new health concern or trend arises, such as an outbreak of a disease or a sudden increase in certain issues, quick identification allows policymakers to react promptly. That's why this thesis focusses on methods for identifying trending or new topics.

Even though there is research on extracting themes and topics from free-text data in questionnaires (Lennon et al., 2021; Nikolenko et al., 2016), there has not yet been any research on real time tracking of topics and trending topic identification in periodic questionnaires. Additionally, there has not been any research comparing the performance of the state-of-the-art methods for this task. Therefore, this thesis compares different methods for real time topic tracking and trending topic identification in periodic healthcare questionnaires.

## 2.2 Literature review

Questionnaires are used for a wide variety of reasons, such as assessing attitudes and opinions, measuring behaviors and gathering demographic information (Hurst & Bird, 2018). Since questionnaires are such a popular research tool, a variety of methods can be and are applied to them. Researchers use methods such as factor analysis, regression analysis, and cluster analysis to analyze quantitative data collected through questionnaires (Chambers & Skinner, 2003).

However, the collection and analysis of open-ended questionnaire responses are less common. This is primarily due to the reliance on human coding, which can be both costly and time-consuming (Yan et al., 2014). However, advances in natural language processing (NLP) and machine learning are

increasingly being utilized to automate and enhance the analysis of open-ended responses. This leads to the use of semi-automated methods (Lieuw et al., 2014), clustering (Lennon et al., 2021) and topic modeling (Nikolenko et al., 2016) to analyze open text data from questionnaires. Even though there is research on topic detection on open text data from questionnaires, there is no research yet on finding the trending topics in repeating questionnaires.

While the detection of trending topics over time has not been performed on questionnaire data yet, it has been done extensively on news, blog and social media data. This can be done in two different ways, by first finding trending tokens and then finding the topics present in those, or by first finding the topics present in the data and then finding the moments when a topic is trending. In the first of these two ways, a number of different methods can be used. These methods include keyword frequency analysis (He & Parker, 2011) and unsupervised term frequency analysis (Tattershall, 2020). Keyword frequency analysis involves tracking the occurrence of specific words or phrases within the dataset to identify spikes that indicate rising interest or concern. Whereas the words tracked over time in keyword frequency analysis are predefined, in unsupervised term frequency analysis, all words are tracked. Among these methods, one of the most widely used is Bursty Term Extraction. This technique identifies terms that experience sudden increases in usage within a specific timeframe, signaling a potentially trending topic. Bursty Term Extraction has thus far been used for detecting trending topics on news (Koike et al., 2013) and twitter data (Xie et al., 2016; Zhang et al., 2017), but not yet on open text questionnaire data.

Additionally, where Bursty Term Extraction finds the trending terms first and extracts the topics from them later, other methods find the topics first and identify the trending topics later. The methods that deal with detecting and tracking the topics in a stream of texts fall within the field of Topic Detection and Tracking (TDT). The initial goal of TDT was “the finding and following of new events in a stream of broadcasting news stories” (Allen et al., 1998). Topic Tracking is the subfield of TDT that deals with the detection of topics and tracking their them over time and can be done both retrospectively and online. In retrospective Topic Tracking the topics in the dataset are found on the whole corpus of texts and then their occurrence over time is tracked retrospectively. In online Topic Tracking, topics are first calculated from a starting set of stories, followed by linking the stories in an incoming stream to one of the existing topics. Within online Topic Tracking two different methods exist, Traditional Topic Tracking (TTT), which follows the framework laid out before, and Adaptive Topic Tracking (ATT). ATT builds on the TTT framework by adding the related stories to the topic model and recalculating the weight distribution based on the new total set of stories. This allows for the evolution and changing of topics over time, making the tracking system more robust. Furthermore, ATT allows for the creation of new topics when a story does not fit well to any of the existing topics.

In the field of ATT, several different methods have been proposed and used over the years. Methods such as the use of semantic classes (Makkonen, 2009) and frequent pattern mining methods (Petkos et al., 2014) have all been used in ATT tasks. However, the most promising state-of-the-art techniques for the task in this research paper make use of either topic chains or Single Pass Clustering (SPC). Topic chains have been used for retrospective topic tracking by Griffiths and Steyvers (2004) and Steyvers et la. (2004) and for online topic tracking by Mei and Zhai (2005) and Kim and Oh (2011). Cantini and MArozzo (2023) created the state-of-the-art topic chain method with the creation of the Length Weighted Topic Chain (LWTC), which creates coherent chains with less noise than prior methods.

Whereas it took a couple of years since TDT’s inception to find the first use of topic chains, single pass methods have been popular in TDT since its inception. There are two primary ways single pass methods have been used for topic tracking. In the first way, each document is compared to the prior documents and threads of similar topics are created. This method was first used in the UMASS

system (Allen et al., 2000) and was later improved by Petrović et al. (2010) for the use of large-scale data streams by the addition of Local Sensitivity Hashing to decrease the total number of comparisons between documents, as a result significantly increasing the processing speed of the system. In the second way, a new document is compared to the cluster centroids of the documents already clustered instead of each individual document. The state-of-the-art SPC method in this domain is developed by Zhe et al. (2011), who added on to the method of Mohd et al. (2011) by adding a sliding time window.

The time series of topic occurrence over time that are provided by the above-mentioned methods allow for the use of other techniques to find the trending topics, namely timeseries analysis. Timeseries analysis techniques such Moving Average and ARIMA can predict future occurrences of topics based on their past occurrences, which allows for comparison of the predicted occurrences with actual occurrences to find trending topics. When the actual occurrence of a topic is much higher than the predicted occurrence, it can be designated as a trending topic. In trending topic detection on time series, Gajtkowski (2022) has used simple moving average with Bollinger Bands to detect trending topics and Mendels (2021) used different forms of moving average and anomaly detection to detect trending topics.

Since there are two distinct ways to find trending topics in open-ended responses, methods from both approaches are used and compared. For the methods that rely on identifying trending tokens and then determining the topics they represent, Bursty Term Extraction is chosen. This method has been effectively applied in various domains (Koike et al., 2013; Zhang et al., 2017) and is reasonably simple to implement. The second approach involves extracting topics first and then identifying which of those topics are trending. This requires two different types of models working together: Topic Tracking models and Trend Identification models.

For the Topic Tracking models, Length Weighted Topic Chain and Single Pass Clustering with a sliding time window are chosen. These models are selected because they are the most promising within Adaptive Topic Tracking, as they represent the most advanced versions of topic chain models and Single Pass Clustering models, respectively. Additionally, their outputs and performances have not been compared yet, which adds value to this thesis.

For the Trend Identification models, three different models are chosen and compared. Moving Average, which was used in both Gajtkowski (2022) and Mendels (2021) to detect trending topics in a topic occurrence time series, showed good effectiveness. Therefore, Moving Average with Bollinger Bands is chosen. Even though ARIMA has not been applied to topic occurrence time series in previous research, it has demonstrated good performance in most time series analysis tasks (Siami-Namini et al., 2018; Sirisha et al., 2022), justifying its inclusion in this thesis. Lastly, Most Occurring Topic Selection, a very naïve solution that simply selects the topics with the highest occurrences, is used as a baseline due to its simplicity and ease of implementation.

## 2.3 Research question

This leads to the research question of this paper:
*How effective are various topic tracking methods (Length Weighted Topic Chain, Single Pass Clustering) and trend identification methods (Most occurring topics selection, Moving Average with Bollinger bands, ARIMA, Bursty Term Extraction) in identifying trending topics in open text fields of questionnaires about health care policies?*

# 3 Data

## 3.1 Data overview

This thesis is conducted in cooperation with the Rijksinstituut voor Volksgezondheid en Milieu (RIVM), the Dutch national healthcare organization, which has provided a dataset with raw text data. The data was generated by an open-ended question asked to members of the general public in a periodic questionnaire about healthcare policies during the COVID-19 pandemic. The questionnaire contained a variety of questions, but for this thesis only the responses to one specific question was analyzed. The specific question is “Heeft u nog iets gemist in deze vragenlijst over corona, dan kunt u dat hieronder schrijven.”, which translates to “If you missed something in this questionnaire about corona, you can write it down below”. Due to privacy reasons the dataset only contains two columns: the responses to the question and from which questionnaire the answer is. The dataset contains 417.505 rows, and each row contains a response from a single individual. In NLP, text fragments that are used as input for a model are described as “documents” and in this dataset each row is considered a “document”. Therefore, in this thesis, the terms “document” and “response” refer to the same object and are used interchangeably. Table 1 gives an example of what the dataset looks like.

*Table 1: A random sample of 5 rows of the dataset. The column response contains all the responses, and the column time step shows at which moment in time the response occurred. There were 15 questionnaires send out at different times, so the time step column goes from 1 till 15, matching each response to its corresponding questionnaire and thus moment in time.*

| ***Response*** | ***Time step*** |
|---|---|
| Meer aandacht voor jongeren. Hier maak ik mij zorgen om. Geef ze de meer vrijheden (zo veel als mogelijk) zodat ze meer invulling aan hun leven kunnen geven. Ook voor kinderen (en eigenlijk ook jongeren) vind ik het van belang dat ze zo snel mogelijk weer kunnen sporten, zowel binnen als buiten. Ik hoop dat hier snel verandering in komt. | 8 |
| Meer maatregelen en het Kabinet moet niet opnieuw te lang wachten anders is de schade vele malen groter | 5 |
| Er moeten altijd boodschappen worden gedaan. Hopelijk krijgen we gauw een prik. | 9 |
| In hoeverre kunnen leerkrachten straks 1,5 m afstand van hun leerlingen in een kleine ruimte houden? Kinderen zitten al weer bovenop elkaar. | 1 |
| Sommige maatregelen zijn niet te begrijpen, waarom moet een zwembad dicht en mag een sportschool open. Theaters etc moeten dicht om verkeersstromen te beperken maar nu loopt men wel massaal in het centrum of in de natuur. | 6 |

## 3.2 Selected data exploration results

The data consist of 417.505 rows, spanning fifteen different collection moments in time. Since each row constitutes a response, we have 417.505 responses in the dataset. The number of responses per collection moment can vary quite drastically, with the most responses for a collection moment being 87.340 and least being 5.349. The responses have character lengths ranging from 0 to 613, with two significant spikes in number of responses per character length. These can be seen in figure 1. Most responses have a character length of either 0 or 28, and a smaller third spike is present at character length 3. The responses with character length 0 are empty and therefore do not hold any value. Of the 103.566 responses with character length 28, 103.160 or roughly 99.6%, are the response ‘Nee, ik heb geen opmerkingen’. This translates into ‘No, I do not have any further remarks.’ So, both the big spikes in character length indicate that they do not have any further comments and these responses

can be removed in the preprocessing. The character distribution after removal of these can be seen in figure 2.

Most of the responses of character length 1 to 4 also indicate a non-response or do not hold any or much value for analysis. For example, all ten most seen responses with character length 3, are a version or spelling mistake of ‘nee’, indicating that they also do not hold any further valuable information. So, all responses of character length lower than 5 are also removed. After removal of all responses meeting these three criteria, the data are reduced from 417.505 rows to 149.447 rows.

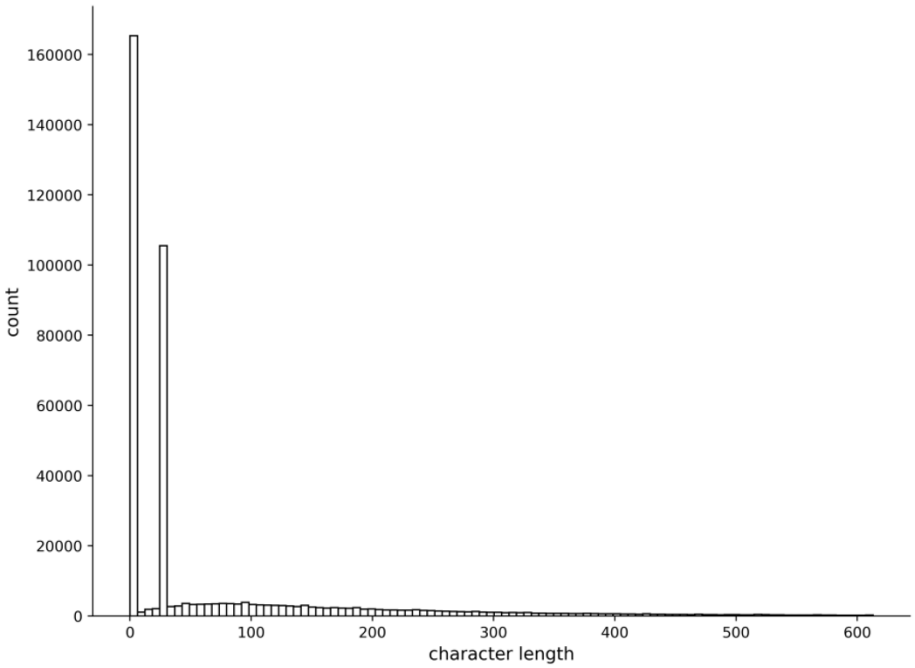


*Figure 1: Character distribution.*

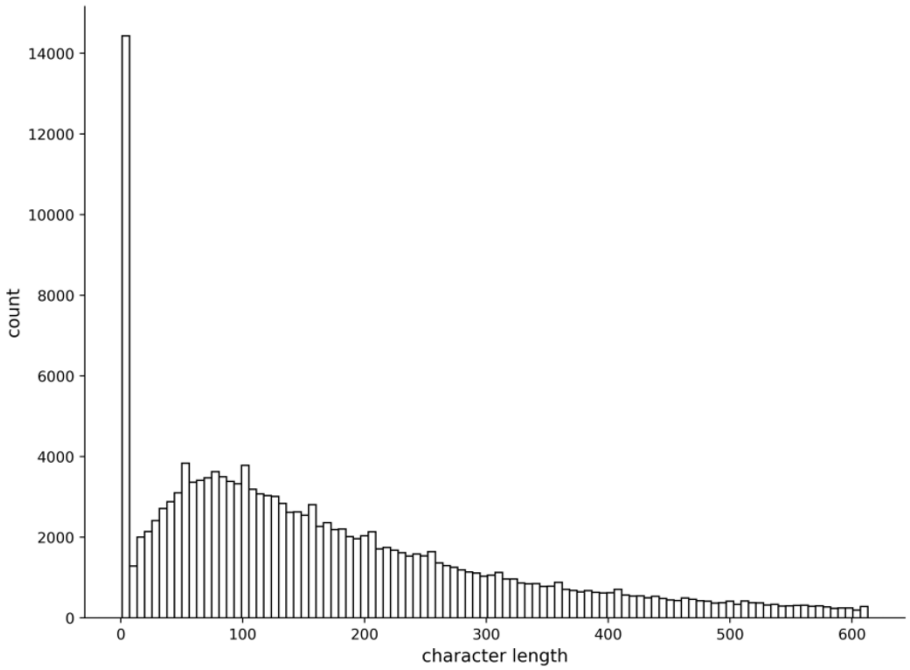

*Figure 2: Character distribution after removal of rows meeting certain criteria.*

In figure 3, a bar plot of the most popular words can be seen. This was made with the data after exclusion of responses with the characteristics mentioned in the prior text and removal of stop words. The most popular words in figure 3 can give an indication of what some of the most important issues in the general public are. These can also show some insight into what topics could be prevalent in the dataset. For example, the two most popular words are 'mensen' and 'maatregelen', or people and measures. So, there could be topics about the effect of COVID-19 on people or about the COVID-19 prevention measures.

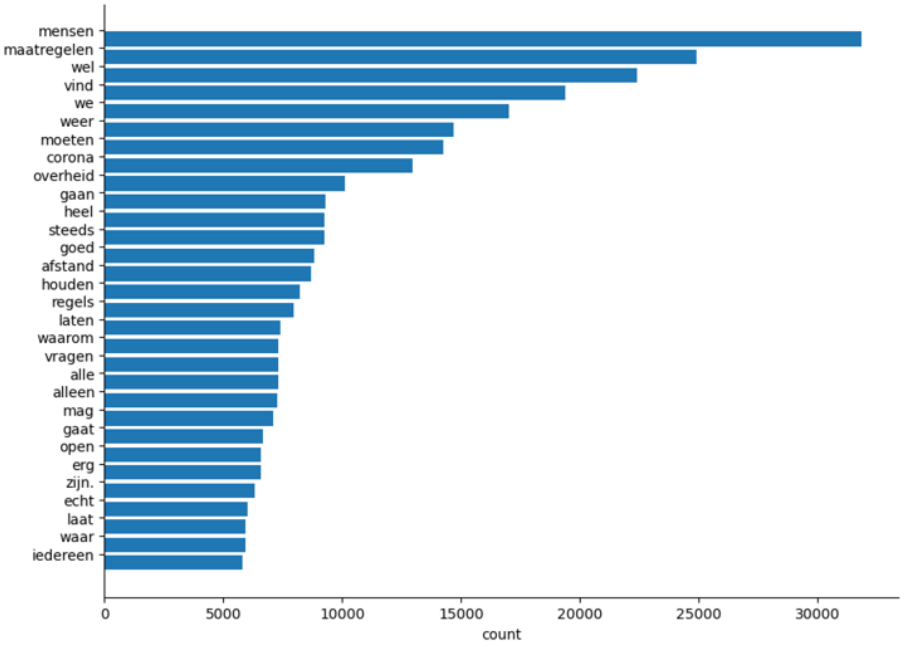


*Figure 3: 30 most popular words after removal of rows meeting certain criteria and stop words.*

## 3.3 Ethical and legal considerations of the data

Due to the sensitive nature of the data and privacy constraints, certain protective measures were put in place. Firstly, the data that were made available by the RIVM had been removed of names and other personal information and included only the raw text of responses. The responses contained in the dataset were only those responses to the one specific question this thesis deals with, and none of the other questions or answers present in the questionnaires were included. Additionally, all the programming with the data was done on internal RIVM servers with the use of a remote portal. This way, none of the data left the RIVM servers at any moment.

# 4 Methods

## 4.1 Overview

The purpose of this thesis is to track topics over time and identify trending topics, and several steps are required to achieve these results. This starts with preprocessing to clean and tokenize the data. Its then split into training, validation, and test datasets. The two Topic Tracking models are trained and fine-tuned, generating topic occurrence timelines. These timelines serve as input for Trend Detection models which are trained and fine-tuned. Additionally, a different Trend Detection model, that uses the preprocessed data instead of the topic occurrence timelines, is used. Finally, both Topic Tracking and Trend Detection methods are used on the test data and their effectiveness is evaluated individually and in combination. The overview of which models are used and in what order can be found in figure 4.

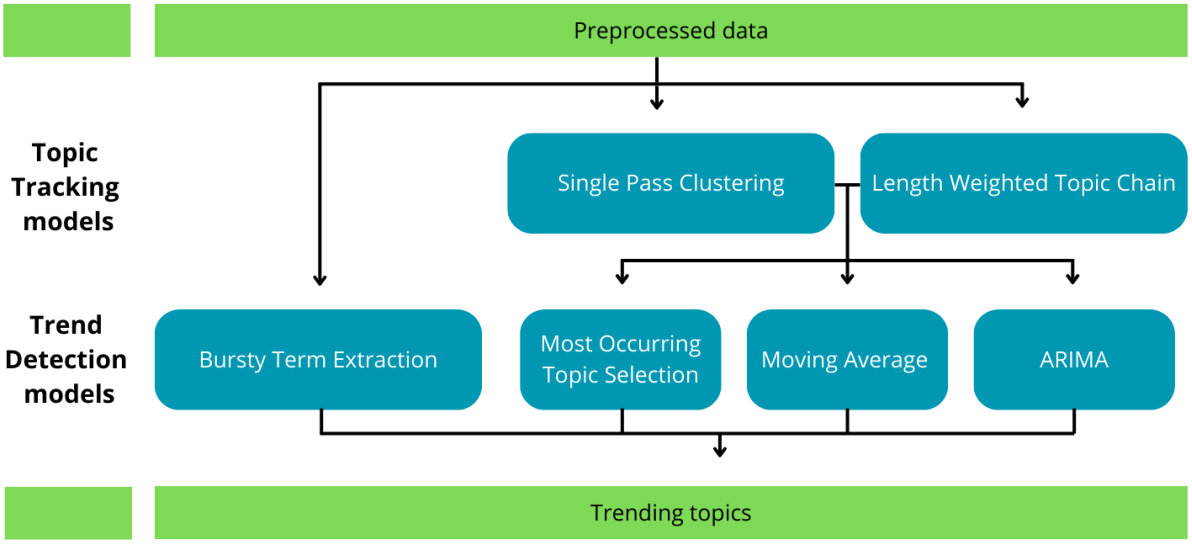


*Figure 4: Overview of which models are used and in what order they are used.*

## 4.2 Preprocessing

To apply the Topic Tracking and Trend Identification models, good quality data are needed as input. To ensure the quality of the input data, the raw data are preprocessed. In exploration of the data, certain problem areas emerged. All the responses containing these three problems were removed, and then the remaining responses were cleaned. In total, the preprocessing is done using the following steps:


*Figure 5: steps in the preprocessing phase*

**Removing empty responses:** This step involves eliminating any responses that do not contain any meaningful text, such as blank or null responses. This is done because empty responses don't contribute to the analysis and might skew the results if included. This is done by using Boolean indexing on a pandas dataframe, discarding the rows that have empty or null responses.

**Removing responses with less than five characters:** This step filters out responses that are too short to provide substantial information, since the responses with very few characters are likely to be noise or irrelevant for analysis. This is done by using Boolean indexing on a pandas dataframe, discarding the rows that have responses of less than five characters.

**Removing responses of "Nee, ik heb geen opmerkingen":** This step involves excluding responses that are exact matches of the phrase, which translates to "No, I have no comments" in English. This is done because it indicates a lack of feedback and is therefore not useful for analysis. Additionally, in the data exploration it was found that a large percentage of the total responses were this exactly, so if not removed they could have a big effect on the analysis. This is done by using Boolean indexing on a pandas dataframe, discarding the rows that are exact matches of the phrase.

**Removing punctuation, digits and excessive whitespace:** This step cleans the text by removing punctuation marks, digits, and unnecessary whitespace, because these can interfere with topic modeling. Regular expressions and string manipulation functions are used to strip out punctuation, digits, and extra whitespace.

**Lowercasing:** This step converts all text to lowercase. Lowercasing ensures consistency in the text data and prevents duplication of words with different cases during analysis. This is done with the lower function for strings in base python.

**Tokenization:** Tokenization is a critical preprocessing step for topic modeling and clustering, as it breaks down the text into individual units (tokens) that serve as the basis for identifying topics within the corpus. The text is tokenized into words at the document level, allowing the topic models to learn the topic distribution per document and clustering models to effectively cluster together similar documents. This is done using the word_tokenize function in the nltk package in python.

**Removing stop words:** Stop words are common words that occur frequently but often carry little semantic meaning. These words are filtered out during text processing to improve the efficiency and accuracy of the topic modeling and clustering models. In this thesis, both Dutch and English stop words are filtered out, given by the Dutch and English stop words lists provided by the nltk package. Additionally, the list of stopwords is extended with a custom list of further Dutch stop words by Diaz (2016). This was done since the Dutch stopwords list of the nltk package is rather limited, with just 101 words being included. The list by Diaz (2016) contains an additional 413 words that carry little semantic meaning, helping increase the quality of the text preprocessing.

In addition to this general preprocessing, some methods might require additional specific preprocessing. These preprocessing steps will be discussed in the sections about the methods in question.

## 4.3 Topic Tracking Models

The topic tracking models identify and follow the main topics in the responses over time. By tracking topic occurrence, these methods create a timeseries that shows how the prominence of each topic changes. Trend identification techniques can then be applied to this timeseries to find trending topics at specific moments in time. The topic tracking models that are used in this thesis are:

- 4.3.1 Length Weighted Topic Chain
- 4.3.2 Single Pass Clustering with a sliding time window

### 4.3.1 Length Weighted Topic Chain (LWTC)

A topic chain is a chain of similar topics in discrete time frames. The data are divided into time slices and the topics in each time slice are found using LDA. Similar topics are then linked together if their cosine similarity exceeds a threshold. This method has been used for retrospective topic tracking by Griffiths and Steyvers (2004) and Steyvers et al. (2004) and for online topic tracking by Mei and Zhai (2005) and Kim and Oh (2011). Cantini and Marozzo (2023) created the state-of-the-art topic chain method with the creation of the Length Weighted Topic Chain (LWTC). The LWTC allows the threshold value to change with the length of the topic chain, the number of links between different time windows. This allows the base threshold value to be lower, making the loss of weaker links smaller, and makes it harder for topics to link to a longer chain. This creates coherent chains with less noise than prior methods.

In this thesis all questionnaires from a single quarter are grouped together since these questionnaires are sent out quarterly. Then LDA topic modeling is performed on each batch separately, and the topics are compared between time steps by the cosine similarity of their word distribution. LDA is further explained in the section under the header *Latent Dirichlet Allocation (LDA)*. When topics between timesteps are similar enough, they are deemed the same and linked together. In LWTC, exponential decay, based on the length of the topic chain, is introduced to improve the performance of the model. The length of the topic chain is defined as the number of links between timesteps. The LWTC algorithm is described in figure 6.

**for each** timestep *i* in timesteps
1. LDA topic modeling is performed on the documents in *i*;
2. **for each** topic in timestep *i;*
   a. calculate cosine similarity with topics in the prior timestep;
   b. adjust the thresholds according to the chain lengths;
   c. **if** similarity ≥ threshold **then** the topics are a topic pair;
   d. **if** similarity < threshold **for** all topics **then** compare with topics from the previous timestep;
   e. This goes on until a topic is linked or the timesteps have run out.
   f. The topics are loaded into an undirected graph network, where the nodes are all the topics, and the edges are the topic pairs.
   g. All the separate networks (connected components) within the graph network that consist of more than one topic, are the topic chains.

end loop.

*Figure 6: Length weighted topic chain.*

The calculation of the threshold is done with the following equation:

$$th_L = th_0 + (1 - e^{-\lambda L})$$

Where:

- $th_L$ is the adjusted threshold value
- $th_0$ is the base threshold value
- $\lambda$ is the threshold adjustment factor
- $L$ is the length of the topic chain, which is the number of links between timesteps

The threshold value determines whether two topics in different time steps are considered a topic pair or not. So, since cosine similarity is used to evaluate the similarity between topics, and the higher the cosine similarity the better matching the topics are, a higher threshold value leads to less topics being linked. The threshold value can be adjusted for the length of the chain by the threshold adjustment factor, which increases the threshold value the longer the chain gets. The higher the threshold adjustment factor, the faster the threshold value rises, which creates shorter topic chains.

***Latent Dirichlet Allocation (LDA)***

Latent Dirichlet Allocation (LDA) is typically a standalone model, but in this case, it is used as a component of the LWTC model rather than as an independent method. LDA is applied separately to the corpus at each timestep to identify the topics present at that specific time. These topics are then compared across different timesteps to create topic chains.

LDA (Blei et al., 2003) is a generative probabilistic model for finding the latent topics in a collection of documents. All the words present in the corpus of documents define a vocabulary. In LDA, each document is assumed to be generated from a probabilistic distribution of topics and it tries to learn this distribution. The probabilistic distribution of topics itself is assumed to be generated from a probabilistic distribution of the words in the vocabulary. LDA tries to learn these distributions via statistical inference, in this case with online variational inference (Hoffman et al., 2010).

In the generative process of LDA, a multinomial distribution *theta_d* over topics is randomly sampled from a Dirichlet, with parameter *alpha,* for each document *d.* To generate each word in *d,* a topic *z_n* is randomly chosen from *theta_d,* and a word *w_n* is generated by random sampling from the topic specific multinomial distribution for *z_n*, which is called *phi_z_n*. The topic distribution *phi_z_n* is obtained by randomly sampling from a Dirichlet with parameter *beta*. From this generative process, the likelihood of a document is obtained:

$$p(w, z, \theta_d, \phi|\alpha, \beta) = \prod_{n=1}^{N} p(w_n|\phi_{z_n}) p(z_n|\theta_d) * p(\theta_d|\alpha) * p(\phi|\beta)$$

It is often sufficient to use fixed values for the parameters alpha and beta, and in following Kim and Oh (2011), these values are fixed at 0.1 for alpha and 0.01 for beta respectively.

### 4.3.2 Single Pass Clustering (SPC) with a sliding time window

SPC is a clustering technique in which each document passes through the system just once. This method requires further preprocessing, as the input for the model is the TF-IDF vector representation of each document. How the TF-IDF representation is created is explained under the header *Term Frequency – Inverse Document Frequency (TF-IDF)*. SPC was first applied to the task of Topic Detection and Tracking by Papka and Allan (1998). The corpus is seen as a stream of documents coming in one after another. Each document in the stream is compared to the existing cluster centroids, and if it falls within a certain threshold it is added to that cluster and the centroid of that cluster is recalculated. If a document does not reach the threshold for similarity with a cluster centroid, it is considered new, and a new cluster is created. The state-of-the-art SPC method in this domain is developed by Zhe et al. (2011), who added on the method of Mohd et al. (2011) by adding a sliding time window. The addition of a sliding time window allows documents to be compared to promising cluster centroids multiple times if they fall just short of reaching the similarity threshold. The document is considered a candidate for a certain cluster, and during the sliding time window, every time the centroid of that cluster is recalculated, the similarity value is calculated again to see if it reaches the threshold now.

Finally, because this allows for the creation of a lot of noisy clusters, which include only 1 or a couple documents, when all documents have passed through, only the clusters containing one percent or more of the total number of documents are kept for analysis. The steps for the SPC model with sliding time window are elaborated in figure 7.

**for each** document *d* in the sequence **loop**
1. find a cluster *c* that maximizes cos(*c, d*);
2. **if** cos(*c, d*) > *t_class* **then** include *d* in *c*;
3. **else if** cos(*c, d*) > *t_candidate* **then** include *d* in candidates for *c;*
    a. **during** window *w;*
        i. **when** *c* gets recalculated, **if** cos(*c, d*) **for** *d* in candidates > *t_class* **then** include *d* in *c;*
4. **else** create a new cluster whose only document is *d*;

end loop.

*t_class* is the similarity threshold value for cluster assignment
*t_candidate* is the similarity threshold value for cluster candidate assignment
*w* is the amount of documents a candidate is kept
a cluster centroid is recalculated every time a new document is assigned to that specific cluster

*Figure 7: Single Pass Clustering algorithm with sliding time window.*

The recalculation of cluster centroids is done with the following equation:

$$T_i = \frac{1}{N} \sum_{j=1}^{N} d_j$$

Where:

- N is the number of stories in one topic cluster.
- $d_j$ is the term vector of every story.

The threshold value determines whether a document is assigned to a cluster or not, and the higher the threshold value, the harder it is for a document to be assigned to a cluster. This leads to more documents not being assigned to a cluster and becoming a cluster of their own, therefore increasing the total number of clusters. This is the case for the class threshold, but the candidate threshold works a little bit different. For the documents that don't reach the class threshold, if they do reach the candidate threshold, they are kept as a candidate for that cluster for the duration of the sliding time window. Therefore, the longer the sliding time window, the more chances a candidate document has to get into a cluster. Additionally, the lower the candidate threshold, the more documents get more chances to get into a cluster. So, both a lower candidate threshold and a longer sliding time window decrease the total number of clusters.

***Term Frequency – Inverse Document Frequency (TF-IDF)***

Term Frequency - Inverse Document Frequency, or TF-IDF, is an often-used statistic to assign the relative importance of a term to a document. It was first used in the 1970's after Inverse Document Frequency (IDF) was proposed in a research paper by Spärck Jones (1972) and was coupled with the existing Term Frequency (TF) metric to create the TF-IDF metric. TF-IDF helps find the words that

occur relatively often in a document compared to the other documents. The functions for TF, IDF and TF-IDF are as follows:

$$tf = \frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d}$$

$$idf = \log\left(\frac{Total\ number\ of\ documents\ N}{Number\ of\ documents\ containing\ term\ t}\right)$$

$$tf - idf = tf * idf$$

## 4.4 Trend Identification models

The output from the Topic Tracking models is a timeseries of topic occurrence. On this timeseries, Trend Identification (TI) methods can be performed to identify trending topics. The created timeseries are quite short, with 15 timesteps, so simple time series analysis techniques are used to detect trending topics. This choice was made because simple timeseries analysis techniques seem to be outperforming more sophisticated techniques on short timeseries (Cruz-Nájera et al., 2022). All but one of the Trend Identification methods used in this thesis use these timeseries as input, but one method, Bursty Term Extraction, works on the preprocessed data itself to find trending topics. As such, this method will be the first one discussed. The four methods that are discussed are the following:

- 4.4.1 Bursty Term Extraction
- 4.4.2 Most Occurring Topic Selection
- 4.4.3 Moving Average
- 4.4.4 ARIMA

### 4.4.1 Bursty Term Extraction (BTE)

BTE is the only trend identification method which is not performed on the time series of topic occurrence that are generated by the topic tracking methods, but instead uses the preprocessed text data itself as input. BTE finds the terms that are bursty, terms that occur significantly more frequently than expected based on its historical baseline. This is done in two steps, selecting bursty terms and removing similar bursty terms. This thesis uses the same steps as Åsblom (2022), which used a combination of methods used by Kunneman and van den Bosch (2014), Li et al. (2012) and Manning et al. (2009).

From the dataset a time sequence of terms, in this case unigrams, is created. Then the expected frequency of each unigram is found using the Gaussian distribution with the following formula (Kunneman & van den Bosch, 2014):

$$E[u|i] = N_i P_s = N_i \times \frac{1}{L} \sum_{i=1}^{L} \frac{f_{u,i}}{N_i}$$

Where:

- $N_i$ is the number of responses received in timestep $i$
- $L$ is the number of time windows containing $u$
- $f_{u,i}$ is the frequency of $u$ in timestep $i$

Following this, a unigram is considered bursty in a timestep if its frequency is higher than the expected frequency (Li et al., 2012). Then, to save time, instead of clustering the bursty terms into bursty topics, similar bursty unigrams are removed if their similarity is too high. This makes sure that unigrams which are lexically similar are not treated as separate. To do this, a similarity score is calculated between each pair of bursty unigrams in a timestep, and unigrams which have a similarity ratio of 90% or higher are seen as equal. This similarity scores makes use of the Levenshtein distance between the two words (Manning et al., 2009). The formula for this similarity score is given below:

$$ratio(s1, s2) = \frac{|s_1| + |s_2| - lev(s1, s2)}{|s_1| + |s_2|}$$

### 4.4.2 Most Occurring Topic Selection

Most Occurring Topic Selection is the simplest TI method used in this thesis. It involves simply selecting the *k* most occurring topics from each of the different time step and selecting them as trending.

### 4.4.3 Moving Average (MA)

A Moving Average (MA) is a simple technique to forecast future values. Several prior data points are chosen, in the case of this thesis three, and the average value of these three data points is calculated. This average value is the value that is the forecast for the current timestep. This forecasted value can then be compared to the actual value. Additionally, for finding trending moments in the timeseries, Bollinger Bands are used on top of the forecasted values. This creates a range of values in which the actual occurrence value could fall and still be considered not trending. The Bollinger Bands are calculated as *k* times the standard deviation around the predicted value. And if the current occurrence value exceeds the upper range of the Bollinger Bands, it is considered trending. The number of standard deviations used for the Bollinger Bands is the parameter which can be tuned here, with a higher value resulting in less trending moments being detected and a lower value resulting in more trending moments being detected.

### 4.4.4 AutoRegressive Integrated Moving Average (ARIMA)

AutoRegressive Integrated Moving Average (ARIMA) builds on the concept of MA and adds more complexity to capture patterns in time series data more effectively for forecasting. It consists of three components: AutoRegressive, which uses past data points to predict future values. Integrated, which involves differencing the data to achieve stationarity. Stationary data have a constant mean and variance over time, which is easier to model. Differencing means subtracting the previous observation from the current observation to remove trends and seasonality. And Moving Average, which smooths out the noise by considering past forecast errors. So, in total, ARIMA uses past values and past forecast errors, after making the data stationary, to forecast a value.

Just like with the Moving Average model, Bollinger Bands are used. This creates a range of values in which the actual occurrence value could fall and still be considered not trending. The Bollinger Bands are calculated as *k* times the standard deviation around the predicted value. And if the current occurrence value exceeds the upper range of the Bollinger Bands, it is considered trending. The number of standard deviations used for the Bollinger Bands is the parameter which can be tuned here, with a higher value resulting in less trending moments being detected and a lower value resulting in more trending moments being detected.

## 4.5 Training and (Hyper)parameter Tuning

To train the models, the data first need to be separated into train, validation and test data. Following this, the parameters for each model can be tuned to find the combination that leads to the best

performance of the model. In the following sections, the splitting into train, validation and test data is discussed first. After this, both the topic tracking models, LWTC and SPC, are discussed separately, since these two models use a different combination of metrics for their parameter tuning. Lastly, all the trend identification models are discussed together, since they all use the same metrics for their parameter tuning.

### 4.5.1 Train, validation and test data

The data are split into a train, validation and test dataset, with 70% of the original data going to the training dataset, 15% going to the validation dataset and 15% going to the test dataset. This results in 104.613 rows for the training dataset and 22.417 rows for both the validation and test datasets. The train, validation, and test splits were performed across the entire timespan rather than by separating different time periods. Since some trend identification methods required previous time steps to function, splitting the data into distinct time sections was not feasible due to the limited number of total time steps available. The topic tracking models are trained on the training dataset and since the methods are unsupervised the best parameter values are found using statistical tests. Afterwards, the models are run with the most promising parameter values on the validation dataset. If the evaluation metrics do not dip a lot, we know that the model is not overfitting to the training data and the best parameters are chosen. The same process is applied to the trend detection models, which are trained on the training dataset and validated on the validation dataset. The final models are used on the test dataset to assess their performance on unseen data.

### 4.5.2 LWTC

For LWTC the parameters which can be tuned are the threshold adjustment factor, the number of topics and the threshold value. In this thesis the threshold adjustment factor is not changed and is set at 0.05, following the paper of Cantini and Marozzo (2023). The number of topics is found through finding the coherence values for the number of topics between 2 and 30 for the responses in the first questionnaire. The number of topics chosen should be as high as possible where the coherence score is not significantly less than the optimal value. This is done because a higher number of topics facilitates the creation of topic pairs and therefore topic chains.

***Coherence score***

To calculate how well the topics are created, a statistical method is used, the coherence score $C_v$ (Syed & Spruit, 2017). In the $C_v$ score the cosine similarity between each topic word and it's topic vector are calculated. The topic vector is created by aggregating all the word vectors of the N most important words for that topic into one big vector. In these word vectors, each cell contains the Normalized Pointwise Mutual Information (NPMI) between that word and all other N most important words for that topic. To calculate the $C_v$ score, the average of all the cosine similarities in the dataset is taken. The function for the $C_v$ score is as follows:

$$C_v = \frac{\sum_{k=1}^{K} \sum_{n=1}^{N} s_{cos}(\vec{w}_{n,k},\ \vec{w}_k^*)}{N \times K}$$

Where:

- N = the number of most important words per topic
- K = the number of topics
- k = topic index
- n = word index in a topic
- $s_{cos}$ = cosine similarity of two vectors
- $\vec{w}_{n,k}$ = vector to represent topic word at index n in topic k
- $\vec{w}_k^*$ = topic vector for topic k

After the number of topics is chosen, the threshold value is found. This is done by finding the threshold value where the number of topic chains is the biggest. And if the number of topic chains is not significantly different for multiple threshold values, the threshold value with the highest number of topic pairs is chosen. To find the optimal threshold value, first the number of topic chains and pairs as a result of the threshold value is calculated on the training data. Then the same is done on the validation data, for the range of threshold values resulting in the best topic chain values on the training data. Then the best combination of threshold values for the validation and training data in that range is chosen as the optimal threshold value.

### 4.5.3 SPC

For SPC, a combination of coherence score and the number of topics found is used to assess the three parameters, class threshold value, candidate threshold value and window size. The exact same coherence score metric $C_v$ is used as in LWTC, so a more detailed description of the metric can be found in the prior section. The additional metric, number of topics found is the number of clusters that is created during an entire run of the SPC algorithm. This metric is added to ensure that the output of the algorithm works well for what it is designed to do. For example, if the highest coherence score is achieved at a threshold value that results in 1.000 clusters, this is not well suited for helping healthcare policy makers assess topical trends, since these topics are probably way too many and too specific. So, a good balance needs to be found between as high a coherence score as possible and a good number of topics for analysis of topical trends.

Because running the SPC algorithm can be very demanding time wise, a grid search on the three parameters was not performed. Instead, first the optimal class threshold value was found without a sliding window or a candidate threshold present. Afterwards, a grid search was performed on the candidate threshold value and window size, with the class threshold value remaining static at the earlier found value. After the optimal combination of parameters on the training data is found, it is assessed whether it reaches similar results on the validation data. If it does not, a range of near optimal parameters in the training phase should be applied to the validation data to assess what the best combination of parameters should be.

### 4.5.4 Trend Identification models

All trend identification models use the same measures for training and parameter tuning, specifically precision and F1-score. Out of the four models, only Bursty Term Extraction (BTE) does not require tuning. For the other three models, the tuning parameters are as follows: for Most Occurring Topic Selection, it is the number of topics selected per timestep; for the Moving Average and ARIMA methods, it is the number of standard deviations used in creating the Bollinger Bands on top of the predicted values.

***Precision, Recall and f1-score***

For the trend identification methods, precision and f1-score are used. To use these measures, the trending topic moments that are found need to be compared to labels. For this, two different labeled timelines are used, which are described in the section under the header *Labeled timelines*. The trend identification methods are applied to the dataset with certain values for the parameter(s) and then the outputs are manually checked against the labeled timelines, which allows for computation of the precision and f1-score metrics. The comparison between the labeled timelines and the trending topic moments is described in the *Labeled timelines* section. How the representations for the topics were made can be found in section 4.5. The formulas for precision and f1-score are given below:

$$Precision = TP/(TP + FP)$$

Where:

- TP is True Positive, the number of instances predicted as true that are actually true
- FP is False Positive, the number of instances predicted as true that are actually false

$$Recall = TP/(TP + FN)$$

Where:
- TP is True Positive, the number of instances predicted as true that are actually true
- FN is False Negative, the number of instances predicted as false that are actually true

The f1-score is defined as the harmonic mean of precision and recall, meaning that it assesses the models ability on both precision and recall. The f1-score for each model is defined as:

$$f1_{score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Where:
- Precision is the ratio of true positive results to the total predicted positives.
- Recall is the ratio of true positive results to the total actual positives.

The f1-score ranges between 0 and 1, with the higher the score the better the model performs.

**Labeled timelines**

The first labeled timeline is a timeline created by manual labeling of part of the dataset by RIVM experts. Prior to this thesis, from all the questionnaires a sample of 1000 responses were labeled using a code book and the most important topics for each questionnaire were found this way. This was done by two researchers, who each labeled 500 responses per questionnaire. For this thesis, only the topics which had an occurrence of more than 25%, so 250 of the 1000 responses, are a trending topic moment. The threshold of 25% was chosen to ensure that only the significant topics were considered as trending topics. This cutoff helps to filter out less relevant or less frequent topics, preventing the timeline from becoming overwhelmed with too many minor topics. Additionally, very broad topics such as 'restrictions', which consist of sub-topics that have a lot fewer mentions are not used in the timeline. This timeline only runs from time step 5 till 14, so not the whole timeline can be compared, but most of it can. The trending topic timeline is given in the table below:

*Table 2: Topical timeline created with the RIVM expert labels.*

| ***Timestep*** | ***Topic*** |
|---|---|
| 5 | Masks |
| 7 | Vaccinations |
| 8 | Vaccinations |
| 9 | Evening curfew and closing of bars, gyms and cultural places |
| | Vaccinations |
| 10 | Dissatisfaction about the government policies |
| | Closing of bars, gyms and cultural places |
| 11 | Corona entrance proof |
| | The restrictions being made more lenient and the pressure this has on the healthcare system |
| 12 | Restrictions and the policy about restrictions |
| 13 | Omikron variant, both the increase in cases because of it and the slowing in hospitalizations |
| | People are done with the restrictions |
| 14 | The restrictions becoming more moderate |

The second timeline is created with the use of the RIVM COVID restrictions timelines (*Tijdlijn van coronamaatregelen 2021.,* n.d.; *Tijdlijn van coronamaatregelen 2020.,* n.d.; *Tijdlijn van coronamaatregelen 2022.,* n.d.). Since these timelines span further than the questionnaires do, all timesteps can be used for comparison to the timeline. For every date of a questionnaire, the two weeks around this date were checked in the online timelines and the restrictions occurring in this range were coupled to that questionnaire. This timeline only includes the restrictions, so it most likely will not include all the topics that are present in the data. The timeline can be seen in the table below:

*Table 3: Topical timeline created with the RIVM covid restrictions timeline.*

| ***Timestep*** | ***Topic*** |
|---|---|
| 1 | Opening of catering industry |
| | Opening of schools |
| 2 | The restrictions becoming more lenient, particularly concerning group sizes |
| 3 | Only six people allowed together indoors |
| 5 | 3 people together indoors |
| | Sport events without audience |
| 6 | Closing of catering industry and public locations |
| | Masks mandatory |
| | Stay at home as much as possible and allowed less visitors |
| 7 | Closing of all non essential businesses |
| 8 | Evening curfew |
| 9 | Slight reduction in restrictions (allowed to exercise in small groups outdoors, going to non-essential store with an appointment) |
| 10 | Corona entrance proof (Coronatoegangsbewijs in Dutch) |
| 12 | Lockdown |
| 13 | Opening of schools and stores, wear a mask |
| 14 | Masks are no longer necessary |

The trending topic moments are checked against the labeled timelines separately, which allows for separate precision, recall and F1-score calculations for the trending topic moments with each timeline. Then these separate scores are averaged over the two labeled timelines to get the average precision, recall and F1-scores for that trending topic moments timeline. The trending topic moments and labeled timeline are checked in the following way:

- The labels for the topics in the trending topic moments and the topics in the labeled timelines are compared manually. They are checked for whether they are about the same topic or not. Some of these are very clear, such as when both the label and the topic in the labeled timeline is “Evening curfew”. Sometimes it is a more subjective call, such as when the label is “Lockdown” and the topics in the labeled timeline are: “Closing of catering industry and public locations”, “Masks mandatory” and “Stay at home as much as possible and allowed less visitors”. In this case, people can come to different conclusion on whether these are about the same topic, but in this case, these were considered to be about the same topics.
- True Positive: If the label for the topic in the trending topic moments and the topics present at that timestep in the labeled timeline match. So, for example, if the label is “Evening curfew” and the topic present in the labeled timeline is “Evening curfew”.
- False Positive: If the label for the topic in the trending topic moments and the topics present at that timestep in the labeled timeline do not match. So, for example, if the label is “Evening curfew” and the topics present in the labeled timeline are “Omikron variant, both the

increase in cases because of it and the decrease in hospitalizations" and "People are done with the restrictions".

- False Negative: If the topic present in the labeled timeline does not match with any of the labels for the topics in the trending topic moments at that timestep. So, for example, if the topic is "Vaccinations" and the labels are "Distance in stores and work" and "Hygiene and safety measures".

## 4.6 Evaluation Metrics

To use the evaluation metrics for this thesis, the words that are most indicative for each topic need to be found. For all the documents in the corpus, a TF-IDF representation is created. Then, for each topic, the mean value of each term is calculated and divided by the mean value of that term over all the topics. This gives a relative TF-IDF value per term, meaning that a term that has a high value in a topic compared to the baseline, is considered more indicative of that topic than a term with a lower relative value. To filter out words that are only used sparingly, such as misspellings, the TF-IDF matrix includes only the terms that occur at least 1000 times in the corpus for the training data and 200 times for the validation and test data, since these datasets are approximately 5 times smaller. Finally, the ten words that have the highest relative TF-IDF values per topic are selected.

The evaluation metric for the Topic Tracking models is word intrusion, a method to quantify the interpretability of topics for humans. The accuracy achieved in the word intrusion task reflects how well the topics are interpreted by humans, providing a measure of the quality of the topic tracking models. This metric is calculated according to the method described by Chang et al. (2009). Here are the steps involved:

1. Five random topics are selected from the total number of topics found.
2. For each of these topics, the five most relevant terms are identified using relative TF-IDF values.
3. An intruder term, which is a term with a high relative TF-IDF value in another topic but a low relative TF-IDF value in the selected topic, is chosen.
4. This intruder term is randomly inserted among the five relevant terms, resulting in a set of six words.
5. For each topic, the participants are presented with the six words and asked to identify the intruder term.
6. The accuracy for each topic is then calculated using the following equation:

$$P_k = \sum_s \mathbb{1}\left(i_{k,s} = \omega_k\right) / S$$

Where:

- $S$ is the total number of subjects
- $i_{k,s}$ is the intruder term selected by subject $s$
- $\omega_k$ is the actual intruder term

7. The final accuracy is the average accuracy achieved between all five topics.

The word intrusion task was conducted through an anonymous survey created with Google Forms and completed by five participants. Each participant assessed the exact same data, which consisted of in total ten topics, five from each topic tracking model. These topics each consisted of six words, which were generated through the method described above. The participants were asked to identify what they thought was the intruder word for each of the topics.

For the evaluation of the Trend Detection models, precision and F1-score are used in the same way as for the parameter tuning. However, instead of on the training and validation dataset, the models are used on the test dataset.

# 5 Results and analysis

## 5.1 Training and parameter tuning

### 5.1.1 LWTC

To find the number of topics used in the LDA model in LWTC, the coherence score is calculated per number of topics used in a LDA model on the first batch of questionnaire responses. This is shown in figure 8. As can be seen from this figure, the optimal coherence score is achieved at 4 topics, but we can see that the coherence score is quite flat at the top and slopes down significantly after around 12 or 13 topics. The LWTC algorithm works better with a few topics, since this allows for more connection chances. So, to strike a balance between an as high as possible coherence score and as many topics as possible, the number of topics to use in the LWTC model is set to eight. In figure 9, the number of topic chains and pairs created based on the different threshold values is plotted for the training data and the validation data. As can be seen, the highest number of topic chains for the training data is created between threshold values 0.725 and 0.825. The optimal threshold value on the validation data is 0.8, while most other threshold values in the range between 0.725 and 0.825 result in significantly less topic chains. This is also the threshold value where there is no big discrepancy between the number of topic chains between the models trained on the training and validation dataset, with the training dataset getting 13 topic chains and the validation dataset getting 12 topic chains. So, 0.8 is chosen as the optimal threshold value.

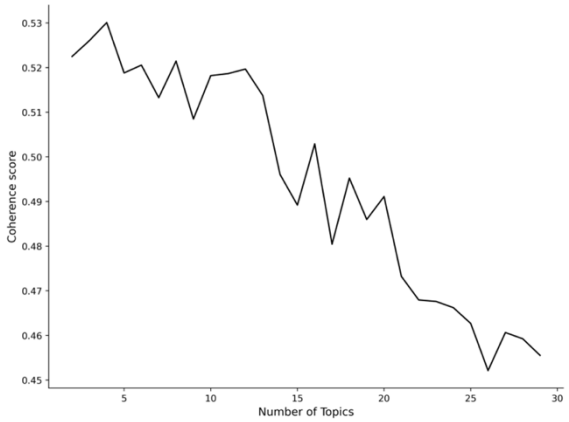


*Figure 8: The coherence score for different number of topics for LDA on the first timestep.*

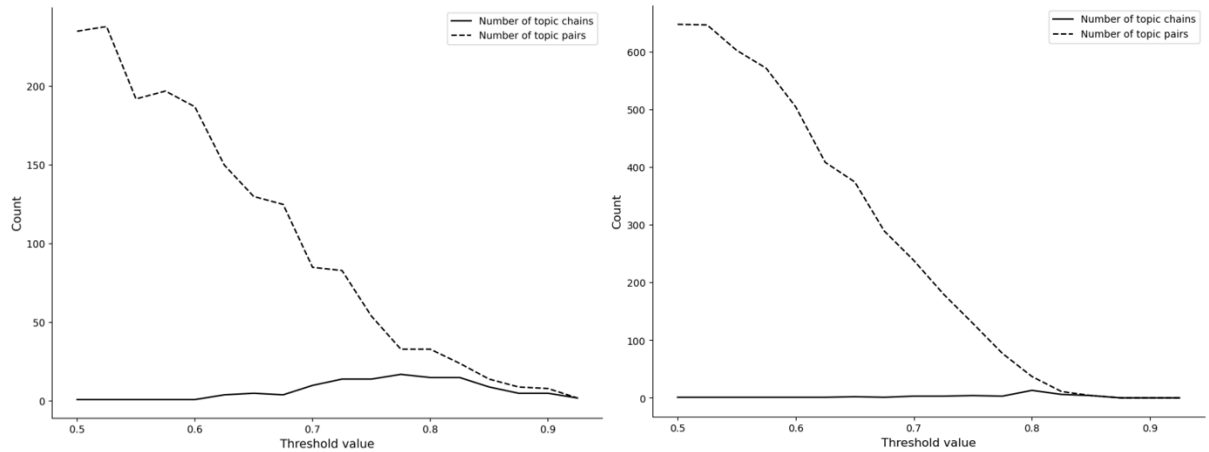


*Figure 9: Number of topic chains and pairs based on base threshold value on the training data (left) and the validation data (right).*

The model is then fitted separately on the training, validation and test data with threshold value 0.8 and the topic timeseries and most relevant terms per topic are calculated. The timeseries and topic terms for the training and validation data are used to train the Trend Identification models, while the timeseries and topic terms for the test data are used to assess the LWTC model performance and as input for the tuned Trend Identification models. The ten words that best describe each topic found in the test data can be seen in table 4.

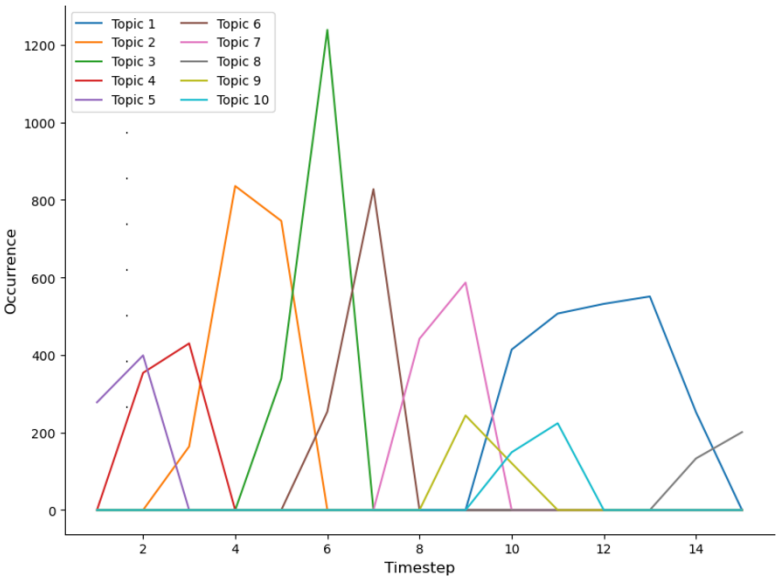


*Figure 10: Timeline created by the best LWTC model using the test dataset.*

*Table 4: List of topics identified by LWTC model on the test data and manually labeled*

| ***Topic id*** | ***Top 10 topic words*** | ***Labels given*** |
|---|---|---|
| 1 | nieuwe, maatschappij, pandemie, besmettingen, samenleve, griep, kabinet, snel, leggen, lange | New wave of covid cases |
| 2 | vragenlijst, vakantie, gevraagd, werk, vraag, vragen, getest, gebruik, afstand, boodschappen | Distance in stores and work |
| 3 | sluiten, musea, theater, app, restaurant, dicht, sporten, strenger, bepaalde, advie | Lockdown |
| 4 | vakantie, familie, gevraagd, vragenlijst, vraag, ivm, meter, hoor, wassen, stress | Family and holiday concerns |
| 5 | vragen, wassen, vragenlijst, vraag, handen, weet, gevolgen, gebruik, gebruik, boodschappen | Hygiene and safety measures |
| 6 | nederland, vaccineren, vaccin, landen, jonge, land, begrijp, laatste, laat, termijn | Vaccination policy |
| 7 | snell, avondklok, vaccineren, vaccinatie, uur, jongeren, hou, begin, termijn, lange | Vaccination speed and curfew |
| 8 | ventilatie, rekene, last, merk, coronamaatregelen, termijn, griep, hele, besmette, lange | Safety measures and long term effects |
| 9 | persoon, begrijpen, bezoek, ontvangen, maatregel, zwaar, belachelijk, avondklok, gezien, onduidelijk | Personal impact of curfew |
| 10 | onbegrijpelijk, controle, horeca, leggen, gebruiken, gevaccineerd, laten, restaurant, keuze, maatschappij | Frustration with regulations |

### 5.1.2 SPC

To find the parameters to use in the SPC model, the optimal class threshold value was found first, followed by a grid search for the candidate threshold value and window size. To find these parameters the number of topics and the coherence score were assessed for different values on both the training and validation dataset. The number of topics and coherence scores as a result of the threshold score are shown in figure 11. As can be seen, there is a peak in number of topics and coherence score for the validation data around 0.0059, whereas these are steady and then drop off for the training data around the same threshold value. So, to achieve the highest number of topics and coherence value, a class threshold value of 0.00595 was chosen.

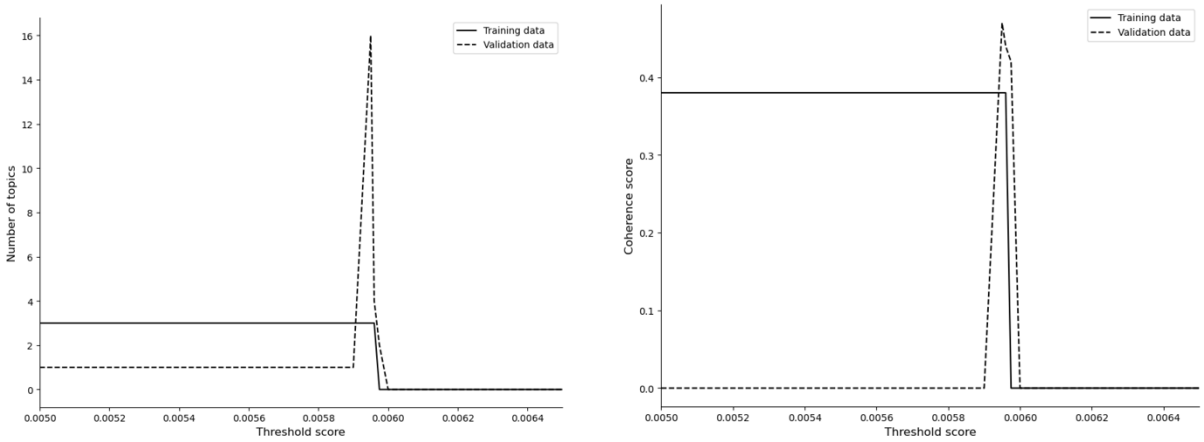


*Figure 11: Number of topics (left) and coherence score (right) based on threshold score for SPC.*

The grid search for candidate threshold value and window size was performed with candidate threshold values 0.005945, 0.00594, 0.005925 and 0.0059 and window sizes 30, 100 and 250. The grid search found that there was no difference in number of topics and coherence score for different window sizes. Meanwhile, candidate threshold values did influence these. The ideal candidate threshold value was found to be at 0.005945, which resulted in the best combination of coherence score and number of topics, with scores of 0.48 and 36 for the validation data and 0.38 and 3 for the training data. All the coherence scores and number of topics are shown in table 5 and 6.

| Candidate threshold value → Window size ↓ | 0.005945 | 0.00594 | 0.005925 | 0.0059 |
|---|---|---|---|---|
| 30 | 0.38 (3) | 0.38 (3) | 0.38 (3) | 0.38 (3) |
| 100 | 0.38 (3) | 0.38 (3) | 0.38 (3) | 0.38 (3) |
| 250 | 0.38 (3) | 0.38 (3) | 0.38 (3) | 0.38 (3) |

*Table 5: Coherence score and number of topics, in the form of Coherence score (Number of topics), for the training data.*

| Candidate threshold value → Window size ↓ | 0.005945 | 0.00594 | 0.005925 | 0.0059 |
|---|---|---|---|---|
| 30 | 0.48 (36) | 0.47 (65) | 0.53 (2) | 0 (1) |
| 100 | 0.48 (36) | 0.47 (65) | 0.53 (2) | 0 (1) |
| 250 | 0.48 (36) | 0.47 (65) | 0.53 (2) | 0 (1) |

*Table 6: Coherence score and number of topics, in the form of Coherence score (Number of topics), for the validation data.*

The model is then fitted separately on the training, validation and test data with class threshold value 0.00595, candidate threshold value 0.005945 and window size 100, and the topic timeseries and

most relevant terms per topic are calculated. The timeseries and topic terms for the training and validation data are used to train the Trend Identification models, while the timeseries and topic terms for the test data are used to assess the LWTC model performance and as input for the tuned Trend Identification models. The ten words that best describe each topic found in the test data can be seen in table 7.

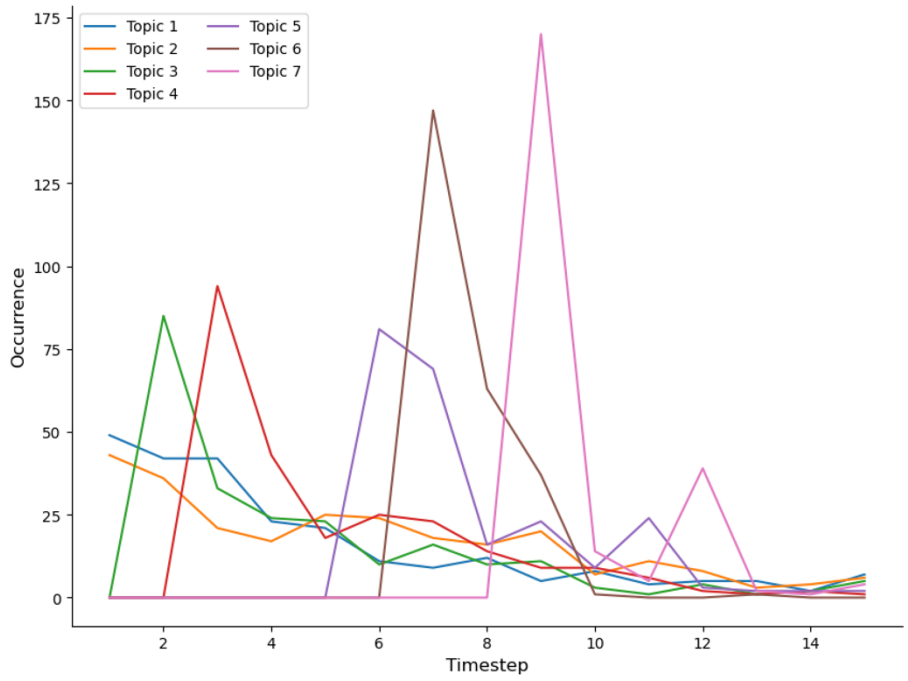


*Figure 12: Timeline created by the best SPC model using the test dataset.*

*Table 7: List of topics identified by SPC model on the test data and manually labeled.*

| ***Topic id*** | ***Top 10 topic words*** | ***Labels given*** |
|---|---|---|
| 1 | wassen, handen, regelmatig, gebruik, boodschappen, plaat, geven, kom, huis, gebruiken | Hygiene and preventive measures |
| 2 | mee, onderzoek, moeite, rekene, gegeven, geval, houd, huis, kom, personeel | Hard to follow regulations at home and work |
| 3 | ventilatie, mis, vragen, rivm, gevraagd, medium, aandacht, cijfer, vraag, ivm | Ventilation and health questions |
| 4 | houden, controle, afstand, moeilijk, virus, lastig, straat, geef, supermarkten, ouder | Social distancing in stores and street |
| 5 | strenger, handhaven, betreft, hard, good, supermarkten, controle, coronamaatregelen, lockdown, langer | Stricter regulations and enforcements |
| 6 | vaccineren, snell, nederland, vaccin, groepen, personeel, landen, laatste, ouderen, onbegrijpelijk | Dutch vaccination speed |
| 7 | open, musea, restaurant, horeca, scholen, theater, veilig, uur, begrijp, onderwijs | Lockdown and reopening |

### 5.1.3 Trend Identification models

The different Trend Identification models are then trained on their respective input, so MOTS, MA and ARIMA on the timelines created with the best LWTC and SPC models and BTE on the responses. In the following sections only the results of the chosen parameters are shared, but the scores for the other parameters that were tested can be found in the appendix.

The MOTS model achieved its best results with the selection of one topic per timestep for both the timelines. On the LWTC timeline this resulted in a precision and F1-score of 0.51 and 0.44 for the training data and 0.54 and 0.54 for the validation data. On the SPC timeline this resulted in a precision and F1-score of 0.47 and 0.46 for the training data and 0.36 and 0.34 for the validation data.

The MA model achieved its best results with the Bollinger Bands at three times the standard deviation on both timelines. On the LWTC timeline this resulted in a precision and F1-score of 0.52 and 0.54 for the training data and 0.56 and 0.59 for the validation data. On the SPC timeline this resulted in a precision and F1-score of 0.38 and 0.17 for the training data and 0.24 and 0.34 for the validation data.

The ARIMA model achieved its best results with the Bollinger Bands at four times the standard deviation on the LWTC timelines and with the Bollinger Bands at three times the standard deviation on the SPC timelines. On the LWTC timeline this resulted in a precision and F1-score of 0.55 and 0.50 for the training data and 0.56 and 0.59 for the validation data. On the SPC timeline this resulted in a precision and F1-score of 0.38 and 0.17 for the training data and 0.32 and 0.41 for the validation data.

## 5.2 Evaluation

### 5.2.1 Word intrusion

To evaluate the clustering model performance, word intrusion scores for the two different methods were calculated. The overall word intrusion score is 0.34. Per method there is quite a big difference, with the LWTC model achieving a word intrusion score of 0.24 and the SPC model achieving a word intrusion score of 0.44. The raw number of participants who identified the right intruder word per topic and the corresponding word intrusion scores can be found in table 8.

There was only one topic where all five participants identified the correct intruder word: "Social distancing in stores and street," created using the SPC method. In contrast, there were two topics where no participants identified the correct intruder word. These two topics both were about regulations, “Stricter regulations and enforcements” from the SPC method and “Frustration with regulations” from the LWTC method, suggesting that topics about regulations are the least well-separated. This could be because topics about regulations are more general and encompass various regulations, making it harder to identify the intruder word. Similarly, topics about vaccinations scored lower in the word intrusion task. The two topics in LWTC, "Vaccination speed and curfew" and "Vaccination policy," received scores of 0.20, while the SPC topic "Dutch vaccination speed" scored 0.40. This lower performance appears to be due to the presence of words that, although relevant, are more general and harder to directly link to the specific topic of vaccinations. Apart from directly related words like "vaccinations" and "vaccine," the most important words often included terms like "Netherlands," "groups," "employees," "countries," and "youth." These terms refer to various demographics or regions and, while relevant to discussions about which demographics receive vaccinations when and what the vaccination policy is in different regions, they make it more challenging to identify the intruder words because they are less specific.

*Table 8: The number of participants who identified the right intruder word per topic and the corresponding word intrusion score. The table is separated into the two methods, LWTC and SPC.*

| **LWTC** | | | **SPC** | | |
|---|---|---|---|---|---|
| *Topic* | *# Good* | *Score* | *Topic* | *# Good* | *Score* |
| *New wave of covid cases* | 3 | 0.60 | *Stricter regulations and enforcements* | 0 | 0.00 |
| *Hygiene and safety measures* | 1 | 0.20 | *Dutch vaccination speed* | 2 | 0.40 |
| *Frustration with regulations* | 0 | 0.00 | *Social distancing in stores and street* | 5 | 1.00 |
| *Vaccination speed and curfew* | 1 | 0.20 | *Hygiene and preventive measures* | 3 | 0.60 |
| *Vaccination policy* | 1 | 0.20 | *Ventilation and health questions* | 1 | 0.20 |
| ***Average*** | 1.2 | 0.24 | ***Average*** | 2.2 | 0.44 |

### 5.2.2 Precision and F1-score

To evaluate the Trend Identification methods, the average precision, recall and f1-scores were computed for MOTS, MA and ARIMA on the LWTC and SPC timelines and for BTE on the raw test data. MOTS and MA were performed with the same parameters on both the LWTC and SPC timelines, being one topic selection for MOTS and three standard deviation Bollinger Bands for MA. On the other hand, the ARIMA models used different parameters for the timelines. For the ARIMA model on the LWTC timeline, three standard deviation Bollinger Bands were used, but on the SPC timeline a model with four standard deviation Bollinger Bands were used. All the average precision, recall and f1-scores for all the different methods are shown in table 9.

On the LWTC timeline, the MA model achieved a precision of 0.63 and f1-score of 0.6 in comparison to scores of 0.57 and 0.58 for the ARIMA and 0.55 and 0.61 for the MOTS methods. On the SPC timeline, the MA and ARIMA models achieved the exact same output and therefore the exact same scores. These two models achieved precision scores of 0.67 and f1-scores of 0.24 in comparison to the MOTS model which achieved scores of 0.39 for both precision and f1-score. The BTE model achieved scores of 0.37 and 0.43 for precision and f1-score.

| | **LWTC** | | | **SPC** | | | **BTE** |
|---|---|---|---|---|---|---|---|
| | MOTS | MA | ARIMA | MOTS | MA | ARIMA | - |
| Average Precision | 0.55 | 0.63 | 0.57 | 0.39 | 0.67 | 0.67 | 0.37 |
| Average Recall | 0.71 | 0.59 | 0.59 | 0.40 | 0.15 | 0.15 | 0.51 |
| Average F1-score | 0.61 | 0.60 | 0.58 | 0.39 | 0.24 | 0.24 | 0.43 |

*Table 9: The average precision, recall and f1-scores for the LWTC, SPC and BTE methods.*

There were some clear similarities between all three models. All three methods identified topics about vaccination policies, curfew, and lockdown, and these were generally considered correctly trending. Both LWTC and BTE found topics focusing on the negative aspects of regulations and policies, such as their adverse impacts and public frustrations, unlike SPC.

However, the most significant difference was between the topics identified by BTE and those identified by SPC and LWTC. SPC and LWTC primarily found general topics, such as vaccinations,

regulations, specific regulatory situations like curfews and lockdowns, and the impacts of these regulations. While BTE also identified some of these general topics, it additionally uncovered more niche topics not found by SPC or LWTC. Examples of these niche topics include the royal family going on vacation, the effects of sex and weight on COVID-19, and testing capacity. BTE also identified several topics about the divide between vaccinated and unvaccinated individuals, highlighting differences in freedom to attend events and visit places between the two groups.

Four different mechanisms explained most of the false positives. Firstly, topics that were not indicative of a specific topic or event often resulted in false positives. An example is the topic "New wave of COVID cases," with the five most important words being “nieuw,” “maatschappij,” “pandemie,” “besmettingen,” and “samenleve.” Secondly, some topics appeared later than those in the labeled timelines, indicating that certain issues remained on people's minds longer. Thirdly, overly general topics, which did not have clear links to specific moments in time, were mostly false positives. Examples include "Hygiene and safety measures" and "Ventilation and health questions." Lastly, overly specific topics, mostly from the BTE method, were identified at the right moments but did not appear in the labeled timelines at all. This occurred because these timelines focused on more pressing and important issues, excluding niche subjects such as “Royal family” and “Religious rule exemptions.”

# 6 Conclusion and Discussion

This thesis aimed to evaluate the effectiveness of various topic tracking and trend identification methods in identifying trending topics in open text fields of questionnaires about healthcare policies. The main findings are as follows.

Single Pass Clustering (SPC) resulted in better-separated topics than Length Weighted Topic Chain (LWTC). However, SPC showed high variability between training, validation, and test datasets, indicating poor generalizability. In contrast, LWTC demonstrated good generalizability, with consistent precision, recall, and f1-scores across datasets.

The Moving Average method was identified as the best for trend identification. Moving Average is followed by ARIMA, then Most Occurring Topic Selection and finally Bursty Term Extraction (BTE) in effectiveness. Interestingly, while SPC produced significantly better-separated topics, LWTC achieved the better results for trend identification, indicating a misalignment between topic modeling performance and trending topic detection performance. This suggests that the quality of topic modeling as measured with word intrusion does not directly translate to effective trend identification. The combination of Topic Tracking methods and Bursty Term Extraction (BTE) also appeared to complement each other, with Topic Tracking methods capturing general trends and BTE identifying more niche topics.

The final results indicated that LWTC with Moving Average is the most effective for analyzing open-ended responses to periodic healthcare policy questionnaires. Despite the high precision of SPC with Moving Average or ARIMA methods, they had very low recall and f1-scores. In contrast, LWTC achieved the highest precision and f1-scores, significantly outperforming both SPC and BTE.

Some of the main limitations of this thesis highlight several areas for future research to enhance the performance and practical usefulness of the methods used. One significant limitation is the subjectivity that is used in the labeling of the topics and the comparison between the trending topics and the labeled timelines. This subjectivity can introduce bias, potentially skewing the results in a certain direction. Furthermore, the lack of expert involvement in this evaluation process is another limitation. Experts were not utilized for the word intrusion tasks, labeling the identified topics or comparing the found trending topics with the labeled timelines. Future research could address this by incorporating domain experts in these tasks.

Additionally, the absence of expert-based evaluation metrics presents another limitation. This thesis did not use a metric that measures the methods' effectiveness based on domain expertise. For example, asking a domain expert to rate the helpfulness of each trending topic during the pandemic on a scale of 1 to 5 could provide valuable insights. Incorporating expert-based evaluation metrics in future studies would help better understand the practical utility of the methods.

Another limitation is the simplified clustering in Bursty Term Extraction (BTE). Due to time constraints, the terms in BTE were not clustered into topics. Instead, duplicates were removed using a similarity measure. This simplification might have affected the accuracy of the topic identification. Future research could look into the performance of BTE if the terms were clustered into topics instead of deleting the duplicates.

Finally, the limitations of the Moving Average and ARIMA methods should be noted. These methods require three historical occurrences to make predictions, preventing them from detecting trending topics in the first three time steps. This limitation reduces the suitability of these methods, and therefore the LWTC with Moving Average method, in scenarios where early trend detection is

critical. Future research could explore modifications to these methods or the use of other methods that can effectively identify trends in the initial time steps.

There are several more promising directions for future research. One potential avenue is the application of LWTC with advanced topic detection methods such as BERTopic or Biterm Topic Model (BTM) instead of LDA. Both BERTopic (Egger & Yu, 2022; Gan et al., 2023) and BTM (Yan et al., 2013) have been shown to significantly outperform LDA in topic detection. It is reasonable to assume that better topic detection would lead to the creation of more accurate topic chains, resulting in improved performance of the LWTC model.

Additionally, since this thesis is the first to directly compare the performance of LWTC and SPC and the first on identifying trending topics in periodic questionnaires, further research could build on these findings. Confirming the results of this thesis through additional studies would provide a deeper understanding of the strengths and limitations of these methods.

Moreover, the disconnect observed between the word intrusion scores and the effectiveness of trending topic identification highlights an intriguing area for future investigation. Exploring this disconnect could lead to the development of more nuanced evaluation metrics and improve the overall accuracy of topic tracking methods.

Overall, the methods studied demonstrate reasonable effectiveness and potential utility for decision-makers, though significant room for improvement remains. The LWTC with Moving Average method emerged as the most effective in analyzing open-ended responses in periodic healthcare questionnaires and could reasonably be used to assist decision-makers at the RIVM. In its current form, this method could be used by RIVM decision-makers to monitor public opinions and concerns through periodic healthcare policy questionnaires.

Additionally, this method could be applied to other forms of short open text data received by the RIVM, such as online portal submissions or transcribed phone calls. Although the application of this method to other data sources should be tested before full implementation, the results from this thesis indicate promising potential. By automating the analysis of such data, this method could either save RIVM decision-makers a significant amount of time or generate new insights, depending on whether the data is currently being manually analyzed or not analyzed at all.

# 7 References


Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study final report.

Allan, J., Lavrenko, V., Malin, D., & Swan, R. (2000, February). Detections, bounds, and timelines: Umass and tdt-3. In *Proceedings of topic detection and tracking workshop* (pp. 167-174). sn.

Åsblom, C. (2022). *Topic Modeling and Detection on a Public Health Case Register* [Bachelor thesis, Vrije Universiteit Amsterdam].

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal Of Machine Learning Research,* 3, 993–1022. https://doi.org/10.5555/944919.944937

Cantini, R., & Marozzo, F. (2023). Topic Detection and Tracking in Social Media Platforms. In *Pervasive Knowledge and Collective Intelligence on Web and Social Media* (pp. 41–56). https://doi.org/10.1007/978-3-031-31469-8_3

Chambers, R. L., & Skinner, C. J. (Eds.). (2003). *Analysis of survey data*. John Wiley & Sons.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems, 22*, 288-296.

Cruz-Nájera, M. A., Treviño-Berrones, M. G., Ponce-Flores, M. P., Terán-Villanueva, J. D., Castán-Rocha, J. A., Ibarra-Martínez, S., Santiago, A., & Laria-Menchaca, J. (2022). Short Time Series Forecasting: Recommended Methods and Techniques. *Symmetry, 14*(6), 1231. https://doi.org/10.3390/sym14061231

Diaz, G. (2016) ”Stopwords Dutch (NL)” [Source code]. https://github. com/stopwords-iso/stopwords-nl.

Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology, 7*, 886498.

Gajtkowski, A. (2022, 7 januari). Predicting FT Trending Topics - FT Product & Technology - Medium. *Medium*. https://medium.com/ft-product-technology/predicting-ft-trending-topics-7eda85ece727

Gan, L., Yang, T., Huang, Y., Yang, B., Luo, Y. Y., Richard, L. W. C., & Guo, D. (2023, October). Experimental comparison of three topic modeling methods with LDA, Top2Vec and BERTopic. In *International Symposium on Artificial Intelligence and Robotics* (pp. 376-391). Singapore: Springer Nature Singapore.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl_1), 5228-5235.

Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. V. (2018). Augmenting Qualitative Text Analysis with Natural Language Processing: Methodological Study. *JMIR. Journal Of Medical Internet Research/Journal Of Medical Internet Research, 20(*6), e231. https://doi.org/10.2196/jmir.9702

He, D., & Parker, D. S. (2011, May). Learning the funding momentum of research projects. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 532-543). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hoffman, M., Bach, F., & Blei, D. (2010). Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23.

Hurst, P., & Bird, S. R. (2018). Questionnaires. In *Research methods in physical activity and health* (pp. 93-101). Routledge.

Kim, D., & Oh, A. (2011, February). Topic chains for understanding a news corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 163-176). Berlin, Heidelberg: Springer Berlin Heidelberg.

Koike, D., Takahashi, Y., Utsuro, T., Yoshioka, M., & Kando, N. (2013). Time Series Topic Modeling and Bursty Topic Detection of Correlated News and Twitter. *International Joint Conference On Natural Language Processing,* 917–921. https://aclanthology.org/I13-1118/

Kunneman, F. A., & van den Bosch, A. P. J. (2014). Event detection in Twitter: A machine-learning approach based on term pivoting. In *Proceedings of the 26th Benelux Conference on Artificial Intelligence* (pp. 65–72).

Lennon, R. P., Fraleigh, R., Van Scoy, L. J., Keshaviah, A., Hu, X. C., Snyder, B. L., Miller, E. L., Calo, W. A., Zgierska, A. E., & Griffin, C. (2021). Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. *Family Medicine And Community Health, 9*(Suppl 1), e001287. https://doi.org/10.1136/fmch-2021-001287

Li, C., Sun, A., & Datta, A. (2012, October). Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 155-164).

Makkonen, J. (2009). Semantic classes in topic detection and tracking.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). Dictionaries and tolerant retrieval [Online]. In *An Introduction to Information Retrieval* (pp. 49–65). Cambridge University Press. https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Mei, Q., & Zhai, C. (2005, August). Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 198-207).

Mendels, O. (2021, 14 januari). Real-Time Time Series Analysis at Scale for Trending Topics Detection - ISE Developer Blog. *ISE Developer Blog*. https://devblogs.microsoft.com/ise/real-time-time-series-analysis-at-scale-for-trending-topics-detection/

Mohd, M., Crestani, F., & Ruthven, I. (2011, June). Construction of topics and clusters in topic detection and tracking tasks. In *2011 International Conference on Semantic Technology and Information Retrieval* (pp. 171-174). IEEE.

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2016). Topic modelling for qualitative studies. *Journal Of Information Science, 43*(1), 88–102. https://doi.org/10.1177/0165551515617393

Papka, R., & Allan, J. (1998). On-line new event detection using single pass clustering. University of Massachusetts, Amherst, 10(290941.290954).

Petkos, G., Papadopoulos, S., Aiello, L., Skraba, R., & Kompatsiaris, Y. (2014, June). A soft frequent pattern mining approach for textual topic detection. In Proceedings of the 4th international conference on web intelligence, mining and semantics (WIMS14) (pp. 1-10).

Petrović, S., Osborne, M., & Lavrenko, V. (2010, June). Streaming first story detection with application to twitter. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 181-189).

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018, December). A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 1394-1401). Ieee.

Sirisha, U. M., Belavagi, M. C., & Attigeri, G. (2022). Profit Prediction Using ARIMA, SARIMA and LSTM Models in Time Series Forecasting: A Comparison. *IEEE Access, 10*, 124715–124727. https://doi.org/10.1109/access.2022.3224938

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation, 28*(1), 11-21.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004, August). Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 306-315).

Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. *2017 IEEE International Conference On Data Science And Advanced Analytics (DSAA)*. https://doi.org/10.1109/dsaa.2017.61

Tattershall, E., Nenadic, G., & Stevens, R. D. (2020). Detecting bursty terms in computer science research. *Scientometrics, 122*(1), 681-699.

*Tijdlijn van coronamaatregelen 2020.* (n.d.). RIVM. https://www.rivm.nl/gedragsonderzoek/tijdlijn-van-coronamaatregelen-2020

*Tijdlijn van coronamaatregelen 2021.* (n.d.). RIVM. https://www.rivm.nl/gedragsonderzoek/tijdlijn-van-coronamaatregelen-2021

*Tijdlijn van coronamaatregelen 2022.* (n.d.). RIVM. https://www.rivm.nl/gedragsonderzoek/tijdlijn-van-coronamaatregelen-2022

Xie, W., Zhu, F., Jiang, J., Lim, E., & Wang, K. (2016). TopicSketch: Real-Time Bursty Topic Detection from Twitter. *IEEE Transactions On Knowledge And Data Engineering, 28*(8), 2216–2229. https://doi.org/10.1109/tkde.2016.2556661

Yan, J. L. S., McCracken, N., & Crowston, K. (2014). Semi-automatic content analysis of qualitative data. *IConference 2014 Proceedings*.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013, May). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1445-1456).

Zhang, T., Zhou, B., Huang, J., Jia, Y., Zhang, B., & Li, Z. (2017). A Refined Method for Detecting Interpretable and Real-Time Bursty Topic in Microblog Stream. In *Lecture notes in computer science* (pp. 3–17). https://doi.org/10.1007/978-3-319-68783-4_1

Zhe, G., Zhe, J., Shoushan, L., Bin, T., Xinxin, N., & Yang, X. (2011, December). An adaptive topic tracking approach based on single-pass clustering with sliding time window. In *Proceedings of 2011 International Conference on Computer Science and Network Technology* (Vol. 2, pp. 1311-1314). IEEE.

# 8 Appendix

## Appendix A: Topic list

Topics from LWTC on training data:

| ***Topic id*** | ***Top 10 topic words*** | ***Labels given*** |
|---|---|---|
| 1 | theater, griep, verplicht, mondkapje, dragen, mondkapjes, stoppen, nieuwe, handen, stop | Public health measures |
| 2 | contact, lockdown, getest, dagen, positief, genoman, zorg, huidige, app, bepaalde | Lockdown and health status |
| 3 | vragenlijst, vragen, getest, klachten, test, vraag, vakantie, mis, werk, testen | Contact tracing and testing |
| 4 | duurt, app, groep, lang, ouder, kinderen, wachten, jongeren, jaar, risico | Age groups and vaccination |
| 5 | vragenlijst, gedrag, normaal, vraag, denken, klachten, rivm, burger, meter, stress | Stress and behavioral response |
| 6 | quarantaine, lange, ggd, positief, app, besmet, horeca, test, onderwijs, controle | Quarantine and positive cases |
| 7 | meter, supermarkt, afstand, merk, anderen, boodschappen, familie, moeilijk, houden, supermarkten | Social distancing in stores |
| 8 | testen, test, dagen, klachten, getest, sind, wachten, vertrouwen, week, hoor | Testing process |
| 9 | snell, vaccineren, vaccinatie, gevaccineerd, vaccin, onduidelijk, gaat, ggd, slecht, wachten | Vaccination speed |
| 10 | avondklok, zwaar, sporten, jongeren, sociale, inmiddel, heen, houdt, merk, houd | Curfew and it's effects |
| 11 | avondklok, uur, duurt, lang, langer, merk, cijfer, vertrouwen, gegeven, steeds | Curfew duration |
| 12 | persoon, ontvangen, bezoek, saman, personen, winkel, open, gezin, restaurant, supermarkten | Restrictions on shops and seeing people |
| 13 | landen, buitenland, lockdown, ziekenhuizen, burger, angst, bevolke, nederland, begin, kabinet | International and domestic covid response |

Topics from LWTC on validation data:

| ***Topic id*** | ***Top 10 topic words*** | ***Labels given*** |
|---|---|---|
| 1 | vragenlijst, dagen, huis, vraag, gevraagd, app, vragen, test, gebruiken, vakantie | Vacation and app and test use |
| 2 | beleid, verplicht, leggen, ziekenhuizen, gevaccineerd, bevolke, laten, ipv, stop, zorg | Strictness of policy |
| 3 | stop, stoppen, mondkapje, maatschappij, besmet, huidige, test, ziek, kwetsbare, quarantaine | Stopping with regulations |
| 4 | vragen, getest, merk, anderen, mis, klachten, mis, meter, vraag, ouderen | Symptoms and behavioral concerns |
| 5 | snell, avondklok, vaccineren, persoon, duurt, vaccinatie, uur, ontvangen, erger, sporten | Vaccination and curfew duration |
| 6 | meter, geval, gevolgen, gevraagd, gevraagde, gezond, handen, vragen, regel, regelmatig | Hygiene and safety measures |
| 7 | strenger, dicht, controle, sind, lockdown, keuze, theater, nederland, vaccin, vaccin | Stricter measures |
| 8 | app, ouder, sluiten, gezin, dagen, contact, school, ten, lockdown | Lockdown and effect on families |
| 9 | covid, voorkoman, alsof, kinderen, hoop, nieuwe, corona, kom, gegeven, sociale | Societal impact |

| | | |
|---|---|---|
| 10 | hoor, vragenlijst, afstand, overheid, meter, deel, neemt, keer, getest, social | Testing and social distancing |
| 11 | gevraagd, verplicht, graag, horeca, begrijpen, leggen, meeste, beleid, waardoor, ten | Public understanding |

Topics from LWTC on test data:

| ***Topic id*** | ***Top 10 topic words*** | ***Labels given*** |
|---|---|---|
| 1 | nieuwe, maatschappij, pandemie, besmettingen, samenleve, griep, kabinet, snel, leggen, lange | New wave of covid cases |
| 2 | vragenlijst, vakantie, gevraagd, werk, vraag, vragen, getest, gebruik, afstand, boodschappen | Distance in stores and work |
| 3 | sluiten, musea, theater, app, restaurant, dicht, sporten, strenger, bepaalde, advie | Lockdown |
| 4 | vakantie, familie, gevraagd, vragenlijst, vraag, ivm, meter, hoor, wassen, stress | Family and holiday concerns |
| 5 | vragen, wassen, vragenlijst, vraag, handen, weet, gevolgen, gebruik, gebruik, boodschappen | Hygiene and safety measures |
| 6 | nederland, vaccineren, vaccin, landen, jonge, land, begrijp, laatste, laat, termijn | Dutch vaccination policy and speed |
| 7 | snell, avondklok, vaccineren, vaccinatie, uur, jongeren, hou, begin, termijn, lange | Speed of vaccination and curfew |
| 8 | ventilatie, rekene, last, merk, coronamaatregelen, termijn, griep, hele, besmette, lange | Safety measures and long term effects |
| 9 | persoon, begrijpen, bezoek, ontvangen, maatregel, zwaar, belachelijk, avondklok, gezien, onduidelijk | Personal impact of curfew |
| 10 | onbegrijpelijk, controle, horeca, leggen, gebruiken, gevaccineerd, laten, restaurant, keuze, maatschappij | Frustration with regulations |

Topics from SPC on training data:

| ***Topic id*** | ***Top 10 topic words*** | ***Labels given*** |
|---|---|---|
| 1 | weten, werk, zwaar, school, kinderen, laten, overheid, maatschappij, denk, mee | Effects on education and families |
| 2 | vragen, bezoek, vragenlijst, familie, kabinet, ontvangen, zien, ouder, persoon, vertrouwen | Family and social interactions |
| 3 | griep, zin, covid, vraag, klachten, voel, groepen, houden, meter, elkaar | Health symptoms and safety measures |

Topics from SPC on validation data:

| ***Topic id*** | ***Top 10 topic words*** | ***Labels given*** |
|---|---|---|
| 1 | weet, test, komt, gevoel, sporten, vraag, bijvoorbeeld, rivm, duidelijk, iet | Testing and clarifications |
| 2 | blijft, voel, blijf, bang, afstand, economie, maanden, weten, contact, ouder | Economic and social impact |
| 3 | vraag, mis, bezoek, ouderen, angst, laatste, gedrag, staat, anderen, weken | Ederly care and behavior |
| 4 | lang, klachten, zoveel, kom, getest, gebruik, begin, zien, vragenlijst, man | Having symptoms for a long time |

| | | |
|---|---|---|
| 5 | werk, zoal, thuis, jongeren, onduidelijk, bezoek, meeste, zit, werken, vragen | Work and youth activities |
| 6 | idee, werkt, denk, gezien, positief, handen, merk, krijg, belangrijk, week | Positive outlook on effectiveness of safety measures |
| 7 | mondkapje, dragen, goede, supermarkt, verplicht, boodschappen, vragen, mene, tijden, jammer | Masks in stores |
| 8 | prima, vragenlijst, vragen, gedrag, samenleve, gebruik, mis, denken, sporten, economie | Effects on sports and economics |
| 9 | weten, economie, corona, samenleve, gedaan, belangrijk, denken, graag, omgeve, informatie | Effect on economy |
| 10 | handen, afstand, groot, situatie, houden, contact, landen, werk, ouderen, kwetsbare | Safety measures |
| 11 | werken, allemaal, thuis, werk, uur, contact, sluiten, ontvangen, ouder, onzin | Working from home |
| 12 | vakantie, terug, buitenland, getest, vraag, meter, huis, nederland, weet, gaan | Travel and testing |
| 13 | handen, normaal, bezoek, staat, duidelijk, huis, mani, blijf, good, gehouden | Guidelines for at home |
| 14 | zeer, mis, vakantie, risico, vragen, week, vraag, vraag, good, nederland | Risks of vacation |
| 15 | meter, gedrag, belangrijk, bijvoorbeeld, maatregel, rivm, voel, afstand, gevolgen, graag | Social distancing |
| 16 | vragen, huis, mis, gedrag, vakantie, mene, rivm, keer, situatie, komt | Home situations |
| 17 | boodschappen, ouderen, komt, jongeren, meter, kwetsbare, supermarkten, afstand, laatste, blijf | Shopping and vulnerable populations |
| 18 | mis, vragenlijst, belangrijk, goede, vraag, gebruik, meter, supermarkt, lastig, man | Social distancing in stores |
| 19 | mondkapjes, dragen, verplicht, stoppen, winkel, ruimte, maatregel, vertrouwen, gebruik, mani | Mask usage and regulations |
| 20 | test, langer, quarantaine, onzin, kom, houden, supermarkten, afstand, testen, kabinet | Testing and quarantaine measures |
| 21 | voldoende, gebruik, mondkapjes, maanden, dragen, supermarkt, veilig, snap, controle, boodschappen | Mask usage in supermarkets |
| 22 | volgen, moeilijk, werkt, lastig, vraag, som, situatie, supermarkt, gebruik, zeggen | Difficulties following regulations while shopping |
| 23 | strenger, controle, handhaven, good, betreft, snel, iet, hoop, regel, verplicht | Stricter measures |
| 24 | open, restaurant, blijven, dicht, scholen, theater, bijvoorbeeld, horeca, winkel, beter | Closing of public places |
| 25 | moment, onduidelijk, duidelijk, som, dingen, advie, lockdown, maatregel, geeft, contact | Uncertainty of measures and lockdown |
| 26 | sluiten, restaurant, theater, dicht, horeca, ruimte, sporten, lopen, druk, supermarkten | Lockdown |
| 27 | theater, quarantaine, positief, aantal, contact, besmet, snel, getest, sluiten, dicht | Testing and quarantaine |
| 28 | coronamaatregelen, overheid, weet, snel, contact, daarnaast, daarnaast, ggd, ggd, ggd | Government measures and information |
| 29 | laat, beter, idee, virus, situatie, regel, tijden, genoman, gevoel, zoveel | Loosening regulations |
| 30 | vaccinatie, landen, vertrouwen, snell, slecht, beleid, nederland, onbegrijpelijk, maanden, vaccineren | Bad vaccination policy Netherlands |
| 31 | burger, avondklok, beleid, daarnaast, graag, hele, totaal, groep, onbegrijpelijk, overheid | Curfew and not understanding |

| | | |
|---|---|---|
| 32 | leven, gehouden, rekene, samenleve, normaal, ipv, vaccin, economie, stoppen, gezondheid | Effects of covid on health and economy |
| 33 | ontvangen, persoon, fijn, bezoek, maatregel, saman, dag, personen, koman, aantal | Rules for visitors |
| 34 | helaas, besmet, bezoek, persoon, vaccinatie, koman, avondklok, lang, controle, regere | Regulations and spreading of the disease |
| 35 | vaccineren, lang, krijgen, snell, slecht, jammer, druk, staat, snel, risico | Vaccination speed |
| 36 | mee, griep, gaat, stop, helemaal, leven, maatschappij, economie, virus, tijd | Negative effects of regulations |
| 37 | hoor, aandacht, aantal, advie, afstand, allemaal, anderen, angst, avondklok, bang | Public fear |

Topics from SPC on test data:

| ***Topic id*** | ***Top 10 topic words*** | ***Labels given*** |
|---|---|---|
| 1 | wassen, handen, regelmatig, gebruik, boodschappen, plaat, geven, kom, huis, gebruiken | Hygiene and preventive measures |
| 2 | mee, onderzoek, moeite, rekene, gegeven, geval, houd, huis, kom, personeel | Hard to follow regulations at home and work |
| 3 | ventilatie, mis, vragen, rivm, gevraagd, medium, aandacht, cijfer, vraag, ivm | Ventilation and health questions |
| 4 | houden, controle, afstand, moeilijk, virus, lastig, straat, geef, supermarkten, ouder | Social distancing hard in stores and street |
| 5 | strenger, handhaven, betreft, hard, good, supermarkten, controle, coronamaatregelen, lockdown, langer | Stricter regulations and enforcements |
| 6 | vaccineren, snell, nederland, vaccin, groepen, personeel, landen, laatste, ouderen, onbegrijpelijk | Dutch vaccination speed |
| 7 | open, musea, restaurant, horeca, scholen, theater, veilig, uur, begrijp, onderwijs | Lockdown and reopening |

Topics from BTE on test data

| ***Timestep*** | ***Top 10 topic words*** | ***Labels given*** |
|---|---|---|
| 1 | tuin | Garden |
| 2 | vliegtuig, wet, demonstratie | Demonstrations |
| 4 | oranje | Royal family |
| 5 | opmerkingen, vrijwilligerswerk, landelijk, testcapaciteit, religieuze | Testing capacity and religious rule exemptions |
| 6 | meld, sportschool, zwembaden, bibliotheek, verbod, kroegen, bioscopen, sauna, feestdagen, zwemle, routekaart, ikea, downloaden, bibliotheken, vuurwerk | Lockdown and closing of public places |
| 7 | voorbereide, januari, begonnen, basisscholen, beginnen, zorgpersoneel, zorgvuldig, start, kerken, beschamend, fesstdagen, gewacht, starten, trage, gestart, essentiële, ingegrepen, gevierd, kerst, vaccinatieprogramma, late, vuurwerk, inentingen | Holidays; Vaccinations; |
| 8 | persoon, rommelig, kapper, logistiek, echtpaar, indicatie, rondje, overgewicht, mannen, perspectief, stemman, verkiezingen, dikke, volgorde, avondklok, vaccineer, leger, vaccinatiebeleid, vaccinatiestrategie, gezondheisraad, britse | Vaccinations policy; Curfew; Elections; Effect of sex and weight on COVID |

| | | |
|---|---|---|
| 9 | Uur, park, afspraak, prikken, snell, tempo, persoon, parken, april, mei, terrassen, langzaam, snelheid, echtpaar, indicatie, avond, uitgenodige, vorig, paspoort, vaart, terras, beurt, jarigen, oproep, schiet, geprikt, dierentuinen, belofte, verloopt, politici, traag, avondklok, vaccineer, leger, klok, rek, astrazenica, sneltesten, vaccinatiebeleid, vaccinatiestrategie, astra | Vaccinations policy and speed; Curfew; Outdoor public places |
| 10 | Evenement, gemeten, meten, dansen, festival, zandvoort, concerten, toegang, voetbalwedstrijden, voetbal, stadion, formule, maten, race, vaccinatiegraad, gevaccineerden, nietgevaccineerden | Events and access; Vaccination status and its effects |
| 11 | Ongevaccineerde, discriminatie, groen, oplopende, tweedele, check, code, vaccinatiegraad, nietgevaccineerde, coronacheck, coronapas, qrcode | Vaccination status and its effects |
| 12 | Musea, cultuur, theater, visie, gooi, variant, sectoren, sector, museum, culturele, ziekenhuisopnames, bioscopen, buurlanden, cultuursector, engeland, kunst, booster, modellen, ikea, zelftesten, boosterprik, boosteren, omikron, omicron | Effects on cultural sector; Omikron and boosters; Comparison to neighboring countries |
| 13 | Carnaval, oorlog, losgelaten | Covid restrictions being let loose |
| 14 | Najaar, opleve | Autumn covid resurgence |
| 15 | Oversterfte, opleve | Covid resurgence and excess mortality |

## Appendix B: Trending topic and labeled timeline comparison results

**SPC Train**

MOTS on SPC train

| | 1a | 1b | 2a | 2b |
|---|---|---|---|---|
| TP | 6 | 7 | 10 | 10 |
| FP | 5 | 8 | 12 | 20 |
| FN | 7 | 7 | 6 | 5 |
| Precision | 0.46 | 0.47 | 0.45 | 0.33 |
| Recall | 0.42 | 0.5 | 0.63 | 0.67 |
| F1-score | 0.44 | 0.48 | 0.53 | 0.44 |
| Avg P | 0.47 | | 0.39 | |
| Avg R | 0.46 | | 0.65 | |
| Avg F1 | 0.46 | | 0.49 | |

MA on SPC train

| | 2a | 2b | 3a | 3b | 4a | 4b |
|---|---|---|---|---|---|---|
| TP | 1 | 3 | 1 | 2 | 0 | 1 |
| FP | 5 | 3 | 3 | 2 | 2 | 1 |
| FN | 12 | 13 | 12 | 13 | 13 | 14 |
| Precision | 0.17 | 0.5 | 0.25 | 0.5 | 0 | 0.5 |
| Recall | 0.08 | 0.19 | 0.08 | 0.13 | 0 | 0.07 |
| F1-score | 0.11 | 0.28 | 0.12 | 0.21 | 0 | 0.12 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Avg P | 0.24 | | 0.38 | | 0.25 | |
| Avg R | 0.14 | | 0.11 | | 0.04 | |
| Avg F1 | 0.20 | | 0.17 | | 0.06 | |

ARIMA on SPC train

| | 2a | 2b | 3a | 3b | 4a | 4b |
|---|---|---|---|---|---|---|
| TP | 1 | 4 | 1 | 2 | 0 | 1 |
| FP | 6 | 3 | 3 | 2 | 2 | 1 |
| FN | 12 | 12 | 12 | 13 | 13 | 14 |
| Precision | 0.14 | 0.57 | 0.25 | 0.5 | 0 | 0.5 |
| Recall | 0.08 | 0.25 | 0.08 | 0.13 | 0 | 0.07 |
| F1-score | 0.10 | 0.35 | 0.12 | 0.21 | 0 | 0.12 |
| Avg P | 0.36 | | 0.38 | | 0.25 | |
| Avg R | 0.17 | | 0.11 | | 0.04 | |
| Avg F1 | 0.23 | | 0.17 | | 0.06 | |

SPC validation

| | MOTS | | MA | | | | ARIMA | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1a | 1b | 3a | 3b | 4a | 4b | 3a | 3b | 4a | 4b |
| TP | 5 | 4 | 6 | 9 | 6 | 9 | 6 | 9 | 6 | 9 |
| FP | 6 | 11 | 23 | 25 | 22 | 24 | 15 | 17 | 15 | 17 |
| FN | 8 | 12 | 7 | 4 | 7 | 4 | 7 | 4 | 7 | 4 |
| Precision | 0.45 | 0.27 | 0.21 | 0.26 | 0.21 | 0.27 | 0.29 | 0.35 | 0.29 | 0.35 |
| Recall | 0.38 | 0.25 | 0.46 | 0.69 | 0.46 | 0.69 | 0.46 | 0.69 | 0.46 | 0.69 |
| F1-score | 0.41 | 0.26 | 0.29 | 0.38 | 0.29 | 0.39 | 0.36 | 0.46 | 0.36 | 0.46 |
| Avg P | 0.36 | | 0.24 | | 0.24 | | 0.32 | | 0.32 | |
| Avg R | 0.32 | | 0.58 | | 0.58 | | 0.58 | | 0.58 | |
| Avg F1 | 0.34 | | 0.34 | | 0.34 | | 0.41 | | 0.41 | |

**LWTC train**

MOTS & LWTC training

| | 1a | 1b | 2a | 2b | 3a | 3b |
|---|---|---|---|---|---|---|
| TP | 4 | 5 | 6 | 11 | 6 | 12 |
| FP | 3 | 6 | 8 | 11 | 11 | 14 |
| FN | 8 | 6 | 6 | 6 | 6 | 5 |
| Precision | 0.57 | 0.45 | 0.42 | 0.5 | 0.35 | 0.46 |
| Recall | 0.33 | 0.45 | 0.5 | 0.65 | 0.5 | 0.71 |
| F1-score | 0.42 | 0.45 | 0.47 | 0.57 | 0.41 | 0.56 |
| Avg P | 0.51 | | 0.46 | | 0.41 | |
| Avg R | 0.39 | | 0.58 | | 0.61 | |
| Avg F1 | 0.44 | | 0.52 | | 0.49 | |

MA & LWTC training

| | 2a | 2b | 3a | 3b | 4a | 4b | 5a | 5b |
|---|---|---|---|---|---|---|---|---|
| TP | 7 | 8 | 7 | 8 | 5 | 6 | 5 | 5 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| FP | 7 | 9 | 6 | 8 | 6 | 7 | 5 | 7 |
| FN | 7 | 7 | 6 | 6 | 8 | 8 | 8 | 9 |
| Precision | 0.5 | 0.47 | 0.54 | 0.5 | 0.45 | 0.46 | 0.5 | 0.42 |
| Recall | 0.5 | 0.53 | 0.54 | 0.57 | 0.38 | 0.43 | 0.38 | 0.36 |
| F1-score | 0.5 | 0.50 | 0.54 | 0.53 | 0.41 | 0.44 | 0.43 | 0.39 |
| Avg P | 0.49 | | 0.52 | | 0.46 | | 0.46 | |
| Avg R | 0.52 | | 0.56 | | 0.41 | | 0.37 | |
| Avg F1 | 0.5 | | 0.54 | | 0.43 | | 0.41 | |

ARIMA & LWTC training

| | 2a | 2b | 3a | 3b | 4a | 4b | 5a | 5b |
|---|---|---|---|---|---|---|---|---|
| TP | 9 | 9 | 7 | 8 | 7 | 6 | 7 | 7 |
| FP | 7 | 10 | 6 | 8 | 4 | 7 | 5 | 7 |
| FN | 6 | 7 | 6 | 6 | 7 | 8 | 7 | 7 |
| Precision | 0.56 | 0.47 | 0.54 | 0.5 | 0.63 | 0.46 | 0.58 | 0.5 |
| Recall | 0.6 | 0.56 | 0.54 | 0.57 | 0.5 | 0.43 | 0.5 | 0.5 |
| F1-score | 0.58 | 0.51 | 0.54 | 0.53 | 0.56 | 0.44 | 0.54 | 0.5 |
| Avg P | 0.52 | | 0.52 | | 0.55 | | 0.54 | |
| Avg R | 0.58 | | 0.56 | | 0.47 | | 0.5 | |
| Avg F1 | 0.55 | | 0.54 | | 0.5 | | 0.52 | |

LWTC validation

| | MOTS | | MA | | ARIMA | |
|---|---|---|---|---|---|---|
| | 1a | 1b | 3a | 3b | 4a | 4b |
| TP | 6 | 8 | 7 | 7 | 7 | 7 |
| FP | 5 | 7 | 5 | 6 | 5 | 6 |
| FN | 5 | 7 | 5 | 4 | 5 | 4 |
| Precision | 0.55 | 0.53 | 0.58 | 0.54 | 0.58 | 0.54 |
| Recall | 0.55 | 0.53 | 0.58 | 0.64 | 0.58 | 0.64 |
| F1-score | 0.55 | 0.53 | 0.58 | 0.59 | 0.58 | 0.59 |
| Avg P | 0.54 | | 0.56 | | 0.56 | |
| Avg R | 0.54 | | 0.61 | | 0.61 | |
| Avg F1 | 0.54 | | 0.59 | | 0.59 | |

**Test results**

BTE results

| | A | B |
|---|---|---|
| TP | 8 | 6 |
| FP | 10 | 14 |
| FN | 5 | 9 |
| Precision | 0.44 | 0.3 |
| Recall | 0.62 | 0.4 |
| F1-score | 0.51 | 0.34 |
| Avg P | 0.37 | |
| Avg R | 0.51 | |

| | | |
|---|---|---|
| Avg F1 | 0.43 | |

LWTC and SPC results

| | LWTC | | | | | | SPC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MO | | MA | | AR | | MO | | MA | | AR | |
| | 1a | 1b | 3a | 3b | 4a | 4b | 1a | 1b | 3a | 3b | 3a | 3b |
| TP | 8 | 6 | 7 | 6 | 7 | 6 | 5 | 5 | 2 | 2 | 2 | 2 |
| FP | 4 | 8 | 3 | 5 | 4 | 6 | 6 | 10 | 1 | 1 | 1 | 1 |
| FN | 4 | 2 | 5 | 4 | 5 | 4 | 7 | 8 | 11 | 12 | 11 | 12 |
| Precision | 0.67 | 0.43 | 0.70 | 0.55 | 0.64 | 0.5 | 0.45 | 0.33 | 0.67 | 0.67 | “ | “ |
| Recall | 0.67 | 0.75 | 0.58 | 0.60 | 0.58 | 0.6 | 0.42 | 0.38 | 0.15 | 0.14 | “ | “ |
| F1-score | 0.67 | 0.55 | 0.63 | 0.57 | 0.61 | 0.55 | 0.43 | 0.35 | 0.25 | 0.23 | “ | “ |
| Avg P | 0.55 | | 0.63 | | 0.57 | | 0.39 | | 0.67 | | “ | |
| Avg R | 0.71 | | 0.59 | | 0.59 | | 0.40 | | 0.15 | | “ | |
| Avg F1 | 0.61 | | 0.60 | | 0.58 | | 0.39 | | 0.24 | | “ | |

## Appendix C: Final trending topic timelines

LWTC final timelines:

| **Time step** | **MOTS** | **MA** | **ARIMA** |
|---|---|---|---|
| 1 | 5 | | |
| 2 | 5 | | |
| 3 | 4 | | |
| 4 | 2 | 2 | 2 |
| 5 | 2 | 3 | 3 |
| 6 | 3 | 3, 6 | 3, 6 |
| 7 | 6 | 6 | 6 |
| 8 | 7 | 7 | 7 |
| 9 | 7 | 9 | 9 |
| 10 | 1 | 1, 10 | 1, 10 |
| 11 | 1 | 10 | 10 |
| 12 | 1 | | |
| 13 | 1 | | |
| 14 | 1 | 8 | 8 |
| 15 | 8 | | 8 |

SPC final timelines:

| **Time step** | **MOTS** | **MA** | **ARIMA** |
|---|---|---|---|
| 1 | 1 | | |
| 2 | 3 | | |

| | | | |
|---|---|---|---|
| 3 | 4 | | |
| 4 | 4 | | |
| 5 | 2 | | |
| 6 | 5 | 5 | 5 |
| 7 | 6 | 6 | 6 |
| 8 | 6 | | |
| 9 | 7 | 7 | 7 |
| 10 | 7 | | |
| 11 | 5 | | |
| 12 | 7 | | |
| 13 | 1 | | |
| 14 | 2 | | |
| 15 | 1 | | |