university of groningen / faculty of science and engineering

The Relation Between COVID-19 Related Public Addresses and Topic Development in Dutch Tweets

Bachelor's Project Thesis

Floris A. van Leeuwen, s3394719, f.a.van.leeuwen@student.rug.nl,
Supervisors: J. Doornkamp, J. K. Spenader

Abstract: Public addresses, given by governmental officials and televised on national senders, are an important tool for COVID-19 related communication between the government and the public. This study examines the relation between public addresses and topic discussion in Dutch tweets from February to December 2020. Topic modelling is performed using the SeaNMF model and data is stripped to limit the computational resources required. The number of resultant topics is decreased by means of manually combining topics into categories. Results reveal sufficient model quality for the identification of long term changes in topic discussion and a high level of variability in shorter time spans. Analysis of the change in topic discussion surrounding public addresses is limited by the variability in topic discussion, yielding inconclusive results. Alternative topic modelling methods are proposed to decrease bias in pre-processing and topic categorization, and the computational load.

1 Introduction

In an attempt to manage the outbreak of the novel coronavirus, the Dutch government imposed their first nationwide restrictions during a press conference on the 12th of March 2020. This and subsequent press conferences received great traction; the 10 most-watched TV moments of 2020 are all either governmental or monarchical public addresses about COVID-19 (Stichting Kijkonderzoek, 2021). In general, the regularly held press conferences were an important tool for the government to maintain support for enforced regulations (van Weelden, 2020).

Twitter has been an important source for news regarding COVID-19, with the Dutch National Institute for Public Health actively distributing the latest information around the COVID-related measures imposed by the government. Short changes in public opinion about politics are often reflected on Twitter (Jungherr et al., 2016). For example, political talk shows are observed to impact the discourse on Twitter during a broadcast (Buschow et al., 2014). These findings indicate that the general popularity of the public addresses might have caused a temporary change in topics discussed on Twitter as well.

Because public addresses were a crucial form of communication between the government and the public during the pandemic, discovering how different public addresses influenced public discourse can provide insight into how the public addresses were perceived by the public. This insight can be helpful in improving governmental communication in times of crisis.

Regarding tweets about COVID-19, Marinov et al. (2020) observed a general shift from global to national concerns in Dutch tweets during the early pandemic. Additionally, Chichirau et al. (2020) indicate that important events might correlate with the number of tweets about some COVID-19 related subtopic. In order to investigate the possible relation between governmental public addresses and public discourse, we performed topic modeling on a dataset of Dutch tweets about COVID-19 from February to December 2020.

The content of public addresses differs. For example, some public address might announce measures impacting social life, while others might impose measures providing financial aid to businesses. Because of this, we expect the change in level of discussion as a result of a press conference to be more prevalent in those topics that are related to the public addresses content. We analyze this change by observing the number of tweets related to a specific topic category on a daily basis and comparing the change around individual public addresses by the overall change.

As Chichirau et al. (2020) noted, many events different from public addresses influenced topic discussion in Italian tweets. Because of this, we hypothesize that a significant change can be observed for some specific topics for some related public addresses, with a smaller difference in change for all public addresses per topic compared to the overall change. For example, the press conference where the possibility of opening international travel is discussed is expected to correlate with a high change in number of tweets regarding travel, while the overall change for this topic regarding all press conferences is expected to be minimal.

Our results indicate the difference in topic discussion is relatively stable over the course of the year, except for the months of February and March, where the difference in topic discussion is significantly different compared to the rest of the year. On a daily basis, the level of variability in topic discussion is high. Although a relation between the date of a public address and topic discussion can be observed for some public addresses, the general effect is too sparse and the variability too large to draw conclusive results.

2 Theoretical Framework

To identify the variety in topics discussed per day, we ran the SeaNMF algorithm following the methodology of Shi et al. (2018). This section describes the process of running the model and model evaluation.

Topic modeling is used to cluster documents into abstract topics by discovering semantic structures in those documents. Recent work into topic modeling has mostly focused on the probabilistic Bayesian-based Larent Dirichlet Allocation (Blei et al., 2003), and the matrix factorization-based Non-negative Matrix Factorization (Lee and Seung, 1999).

Although these methods have been successful in extracting features from long-form documents, traditional methods for topic modeling often perform badly on short documents like Twitter data (Shi et al., 2018). The semantics-assisted NMF model proposed by Shi et al. (2018) is observed to effectively solve this problem. Early tests on about 15 categories of tweets with 2500 to 3000 tweets per topic show the SeaNMF algorithm significantly outperforms the regular NMF algorithm on tweets. Furthermore, the algorithm has been observed to be successful on both Dutch and Italian tweets regarding the COVID-19 pandemic (Marinov et al., 2020; Chichirau et al., 2020).

This section will first go over how NMF can extract topics from a set of documents. After that, the problems that arise when the NMF algorithm is used on short documents are stated and the semantics assisted solution proposed by Shi et al. (2018) is illustrated.

2.1 Traditional NMF

The NMF algorithm is an unsupervised algorithm that utilizes document-level word co-occurrence to extract topics from a set of documents. The model transforms a bag of words representation of all documents into a list of documents and their topics and a list of topics and their keywords. This is done by factorizing a high dimensional non-negative input matrix A into two lower dimensional non-negative matrices H and W, such that $A \approx WH$. The matrix A then consists of document-keyword pairs derived from the corpus.

The matrix W contains the inferred keyword-topic correlations and the other resultant matrix H contains the inferred document-topic correlations. In other words, W contains the list of topics and H contains the list of documents by topics. Because A is factorized into two matrices, a hypervariable K is introduced to dictate the dimensions of the resultant matrices. The value of K represents the number of topics, so that for matrix A of size m distinct keywords by n documents, the dimensions are $A_{n\times m} = H_{n\times k} W^T_{m\times k}$.

W and H are converged by iterating over a goal function. Many specific different varieties of goal functions for the NMF algorithm have been defined (Cichocki et al., 2009). In general, the difference between A and the reconstruction of A as the product of W and H is determined as the loss of the algorithm. Convergence is terminated when some pre-defined minimum loss is obtained. An example of a customary objective function provided by Shi et al. (2018) is displayed in Equation 2.1. Here, the minimum difference between A and WH^T is determined using the Euclidean distance between the two matrices. A visual representation of the NMF model is presented in Figure 2.1.

$$\min_{W,H \leq 0} ||A - WH^T||_F^2 \tag{2.1}$$

2.2 Problem statement; NMF and short documents

The NMF algorithm uses document-level word co-occurrence to discover topics in a set of documents. Since a small number of keywords per document results in a low number of word co-occurrences, the semantic patterns the NMF model utilizes are sparse. Because of this, the performance of the NMF algorithm depends on document keyword length (Hong and Davison, 2010; Zhao et al., 2011).

Tweets are inherently short documents because of the maximum of 280 characters per tweet. Furthermore, pre-processing steps, including stop word removal, further decrease the number of available keywords. The removal of terms is further elaborated in Section 3.1. Because of this, the number of keywords provided by tweets is insufficient to reliably extract topics from Twitter corpora (Shi et al., 2018).

The techniques used to tackle this problem can be categorized into three groups. Hong and Davison (2010) propose to bundle tweets by their meta-data to increase the length of the documents. Another strategy is to integrate semantic word correlation into the algorithm (Sridhar, 2015; Xun et al., 2016). For example, Xun et al. (2016) used word vector embedding to increase contextual information. Lastly, the SymNMF algorithm proposed by Kuang et al. (2012) uses a similarity measure between data points as an input, as opposed to a data matrix used by the regular NMF algorithm. This performs well on discovering nonlinear relationships in data, but for the clustering use case of topic modeling, it is found to yield unintuitive results (Shi et al., 2018).

Shi et al. (2018) proposes a different semantics-assisted adaptation of the NMF algorithm. Here, in addition to the document term matrix, a matrix that represents the semantic relationship between keywords and contexts is generated. The difference here with the methodology of the SymNMF model, is that no external sources are used to obtain semantic information. This is expected to yield better results by reducing the amount of noise and bias introduced (Shi et al., 2018).

2.3 Semantics-assisted NMF

Shi et al. (2018) proposes a solution for this problem of data sparsity in short documents called the semantics-assisted NMF (SeaNMF) model. In this model, semantic relationships between keywords are discovered by introducing contexts, where every context represents a group of semantically related keywords. The general difference between the NMF and the SeaNMF model is the introduction of a new matrix S and the updating of the objective function to account for this. This new matrix S provides information about semantic relationships between words from different documents, effectively decreasing the reliance of the NMF algorithm on document size. The factor of the combinations of matrices A and S result in a new matrix W_c in addition to W and H, revealing the inferred relationship between keywords and contexts.

S is determined by means of neural word embedding. Shi et al. (2018) uses a skip-gram model to determine a unigram distribution for each sampling context. Because the number of keywords in a single document is small, the sliding window size for each keyword is equal to the length of the document. Therefore, the number of windows is equal to the number of documents. This is based on the assumption that semantically related words appear more often together within documents than non semantically related words. The semantic relation derived from word co-occurrences is used to generate contexts as abstract longer form documents. The unigram distribution for a context represents the prevalance of the context terms in the corpus. It is determined by Equation 2.2, where $\#(c_j)$ is the number of occurrences of context c_j and γ is a smoothing factor.

$$p(c_j) = \frac{\#(c_j)^\gamma}{\sum_{c_j \in V} \#(c_j)^\gamma} \tag{2.2}$$

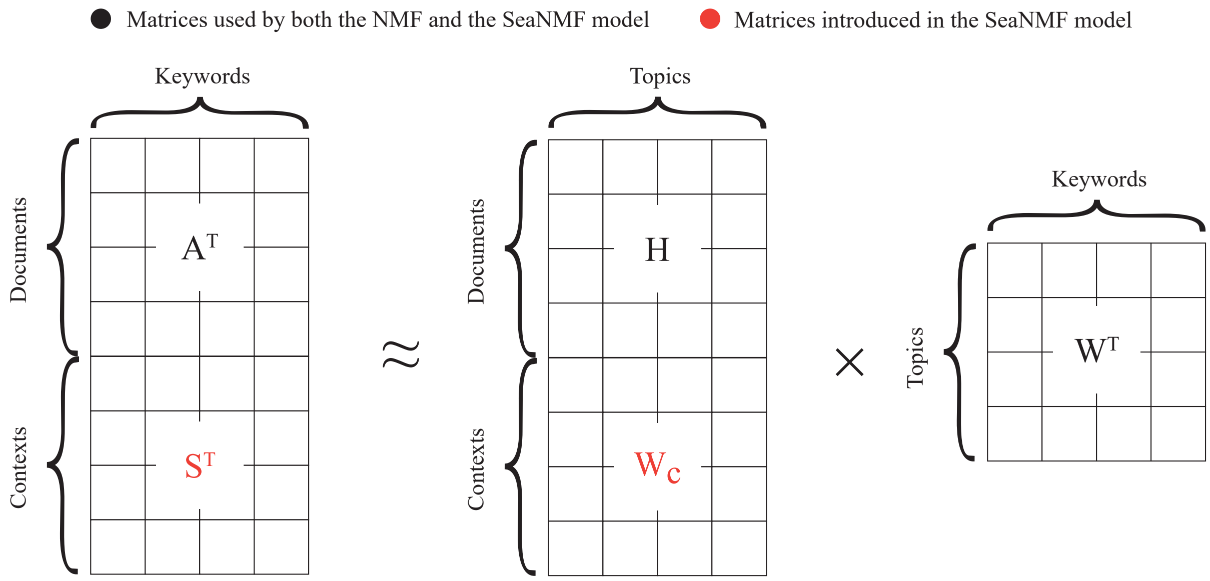

Figure 2.1: A visual representation of the objective of the NMF and the SeaNMF model. A **is the input document-keyword matrix;** S **introduces semantic information into the input;** H **represents the documents in terms of topics;** W_c **represents the contexts in terms of topics;** W **represents the topics in terms of keywords.**

The correlation between a word and a context S_{ij} is described by Equation 2.3. Here, the number of word-context pairs in the corpus $(\#(w_i, c_j))$ is divided by the total number of word occurrences $(\#(w_i))$ and the distribution for the context $(p(c_j))$. This means that every term-context combination is assigned some correlation score, where semantically related terms together have a higher level of correlation for some context. Ubiquitous terms will generally have some correlation for most contexts, with a low average correlation level. Note that every negative value in S is set to zero, in accordance with the constraints of the traditional NMF algorithm.

$$S_{ij} = \left[log \left(\frac{\#(w_i, c_j)}{\#(w_i) \times p(c_j)} \right) \right]_+ \tag{2.3}$$

The resulting objective function for the SeaNMF model proposed by Shi et al. (2018) results from $S \approx WW_c^T$ and the traditional objective function (Equation 2.1). In Equation 2.4, the objective function is presented as the minimum difference between the combination of A and S and the combination of the derived matrices H, W_c and W. ψ is a penalty function described by Shi et al. (2018) and α is a scaling factor described in Section 3.2.2. A visual representation of the difference between the the traditional NMF's and the SeaNMF's objective function is presented in Figure 2.1.

$$\min_{W, W_c, H \geq 0} \left\| \begin{pmatrix} A^T \\ \sqrt{\alpha} S^T \end{pmatrix} - \begin{pmatrix} H \\ \sqrt{\alpha} W_c \end{pmatrix} W^T \right\|_F^2 + \psi(W, W_c, H) \tag{2.4}$$

3 Methods

3.1 Data pre-processing

Topic modeling is performed on a dataset of 5, 491, 551 Dutch tweets about COVID-19 from February to December 2020 (Caselli and Basile, 2020). Trending COVID-19 related hashtags were extracted from a trend website on a daily basis. Using this, a set of the most common COVID-19 related hashtags was defined, which in turn were used to scrape the tweets.

Pre-processing the document contents is an essential part of the SeaNMF algorithm because of its dependence on word co-occurrence. In general, the pre-processing steps are taken to remove noise from the input data and reduce the initial dataset size, thus reducing the algorithm's computational load. The pre-processing steps were the following respectively:

1. The keywords used for scraping are removed. Since at least one of these keywords is present in every document and we are interested in the difference in topics within the general COVID-19 topic, removing these prevents overfitting on these keywords. A tweet containing many terms regarding China is displayed below. We observe that many terms besides the scraping keywords represent the topic of China. Because this is what is generally observed by us from the data, the prevention of overfitting by keyword removal is judged to be more beneficial than the inclusion of the scraping keywords in the vocabulary.

   > "Een tijdje terug kwam dit bericht uit #Wuhan (#China) van een gewone vrouw. Zij lucht haar hart over het corrupte #CCP regime in #Beijing. #WuhanVirus #ChinaVirus **#coronavirus #COVID19 #Covid_19 #CoronavirusOutbreak** #coronaviruswuhan **#coronaviruschina** https://t.co/4ZYR91ZuT3"

   – Tweet ID: 1232376058407682000. The removed keywords used for scraping are displayed in **bold**.

2. Hyperlinks, mentions (the tagging of other users in a tweet), numerals, and all nonalphabetic characters are removed because these mostly do not contain any semantically significant information.

3. Stop words * and closed-class words † are removed. The removal of frequent and trivial tokens ensures only the most relevant keywords are included in the generated vocabulary.

4. The resulting set of terms is lemmatized † and lowercased. This enables the model to pick up on the same words that are present in a different grammatical form or different part of speech.

5. Lastly, duplicate documents are removed, because a high number of duplicate tweets decreases the validity of the assumption that semantically related words occur together more frequently in a document.

3.1.1 Data stripping

The SeaNMF algorithm is not suited for large corpora, because of the exponential growth of matrix A and S by the input matrix size, which in turn results in slow convergence as well (Lin and Boutros, 2020). Steps are taken to decrease the computational load of the algorithm. The space complexity of the model is displayed in Equation 3.1, where m and n are the number of distinct keywords and the number of documents respectively.

$$\mathcal{O}(m^2n^2) \tag{3.1}$$

We observed that the memory usage that results from the number of documents in the dataset is significantly more than the processing resources we have access to for this project. This problem was also encountered by Marinov et al. (2020) and Chichirau et al. (2020), who proceed by parting the data into months and evaluating the model over each month individually. This comes with its own problems, however, because the resulting topics are different for each month and were manually grouped together subsequently. This drastically increases the number of assumptions introduced.

The space complexity (Equation 3.1) shows that the required memory can be reduced by both reducing the number of documents and reducing the number of distinct keywords. The pre-processing steps discussed before already decrease the number of distinct keywords significantly. Another factor in reducing the number of keywords is the hyperparameter n_t, which dictates how often a term should be present in the corpus to be included in the input vocabulary. Increasing this value to decrease the number of distinct keywords results in significantly poorer performance. Shi et al. (2018), Marinov et al. (2020), and Chichirau et al. (2020) show that including a large vocabulary size is essential in retrieving satisfactory results from the model. For this reason, memory usage is not decreased by means of decreasing the number of keywords.

Shi et al. (2018) shows that only a small dataset is needed to extract relevant topics from a list

*For this, the Python `stop-words 2018.7.23` library was used.

†For this, the NLP `Spacy` library was used.

of documents. Therefore, documents are removed from the corpus after the initial pre-processing steps in order to drastically decrease the matrix sizes. Because the goal is to inspect differences in topics discussed on a timeline, we want to balance the dataset so no month has a prior bias in training due to differences in size. The month with the lowest number of tweets is February, with 42836 unique tweets, therefore tweets are removed randomly from every month until their individual size consists of 42836 documents. This ensures the data is distributed evenly per month while decreasing the number of data points to be evaluated, thus reducing computational load.

After all pre-processing, the total number of tweets was decreased from $5,491,551$ to $471,196$. This is found to sufficiently decrease the memory required to run the SeaNMF algorithm.

3.2 Tuning

Three hyperparameters are tuned to our specific dataset and evaluated using two distinct evaluation metrics. This section first describes the evaluation metrics, which are then used to describe the tuning of the algorithm.

3.2.1 Evaluation metrics

Two evaluation metrics are used. The Pointwise Mutual Information (PMI) score indicates the level to which terms within topics correlate with each other and the Topic Diversity (TD) score indicates the uniqueness of the terms in a topic compared to other topics.

The PMI score describes the association between two terms as the probability of the two terms occurring within a document compared to the probabilities of the two terms occurring independently (Equation 3.2).

$$PMI = log\left(\frac{P(x,y)}{P(x) \times P(y)}\right) \tag{3.2}$$

A PMI score of 0 then means the two keywords appear as frequently as expected, a negative score means the two keywords appear together less frequently than expected, and a positive score means the two keywords appear together more frequently than expected. The latter indicates a possible semantic correlation between the two keywords. The mean PMI score of all top n_k keywords within a topic with all other top n_k keywords within that document is used as an evaluation metric to judge the topics validity. For n_k, a value of 10 is used. This is in line with the value used by Shi et al. (2018), so this allows us to properly compare our scores to theirs. Furthermore, while increasing the value of n_k we observe no notable relative difference in PMI score between topics, thus not retrieving additional information.

The Topic Diversity metric (TD) assigns a uniqueness score to a topic using the number of unique top n_k keywords for that topic. To determine the TD we use an adaptation of the approach used by Chichirau et al. (2020). Equation 3.3 describes the Topic Diversity metric TD for each topic i, so that a TD of 1 means the top keywords for a topic are completely unique over all other topics. A TD of 0 occurs when all top keywords in the topic appear in all top keywords in the other topics, effectively meaning there is a high level of similarity between all topics.

$$TD_i = \left(1 - \frac{\sum_{j \neq i} u_{ij}}{n_k * K}\right)^2 \tag{3.3}$$

In Equation 3.3, u_{ij} is the sum of the total number of unique occurrences of keywords in topic i, compared to topic j, n_k is the number of top keywords to evaluate per topic and K is the total number of topics. The result is squared to increase the difference in high values of TD for readability purposes. For determining the TD, a value of 30 is used for n_k. This value is determined by increasing n_k to reduce the number of topics with a TD of 1, thus yielding more information.

3.2.2 Tuning Process

Three tuning parameters are tuned to maximize both the PMI and TD. This subsection goes over them one by one including their desired values, after which the tuning process is described.

Parameter n_t describes the minimum term frequency for a term to be included in the vocabulary used for creating matrix A. Terms that only appear a single to a few times in the vocabulary provide a negligible amount of semantic information, because of the low number of co-occurences for that term. This means that for n_t, the value should be tuned to only exclude the most rare terms. As described in Section 3.1, a large vocabulary size is essential for the model to return satisfactory results. Therefore, while tuning we aim to keep n_t low, while making sure we stay within the computational resources available to us.

Another tuning parameter is K, which describes the number of topics to be retrieved from the model. The number of sound COVID-19 related subtopics retrievable from the dataset is not known beforehand, therefore we tune this value mainly by observing the TD and PMI score. Allthough, it should be noted that a low value for K is generally prefered over a higher value. This is because a large number of distinct topics will result in more assumptions made while categorizing the topics. This is elaborated in Section 3.4.

The last tuning parameter α is the scaling factor of the semantic correlation matrices (S and W_c, Equation 2.4). α effects the size of the effect of the introduced semantic information by the SeaNMF model.

Using the information about n_t, K, and α mentioned above, we start the tuning process with a value of 1 for α, as this is the value used by Shi et al. (2018). Because of the co-dependence of n_t and K on TD and PMI described in Section 3.2.3, we then tune these parameters simultaneously in blocks of different values for n_t. After optimal values for n_t and K are found, α is tuned.

3.2.3 Tuning Results

The results of the tuning process for the parameters K and n_t are presented in Figure 3.1. A large n_t results in the lowest PMI and TD. This is because a larger n_t results in a smaller vocabulary and therefore a smaller probability of keywords being unique within a topic. The probability of keywords being unique within a topic also decreases with a smaller K, because the number of topics compared decreases. This indicates that for maximizing TD, n_t and K are co-dependent. Therefore, we tune for multiple values of K within values of n_t.

Decreasing n_t also results in a higher PMI. As the size of the vocabulary is larger, the chance of the corpus containing patterns of word co-occurrence increases. A larger value of K is also observed to correlate with a larger PMI, however, at some n_t dependant threshold this effect is observed to be negligible. This is because a low number of topics can result in less semantically similar words being combined within a topic's top keywords, reducing the PMI score.

Combining the results presented in Figure 3.1, the desire to keep n_t low for computation purposes and (Section 3.1.1), and the desire to keep K low for evaluation purposes (Section 3.4), a value of 50 for K and 100 for n_t is used for training the model. After 50, the effect of increasing K becomes trivial, and lowering n_t even more would cause the computational resources to overflow.

Lastly, α is tuned. We observe an increase in PMI and TD score while increasing α to 9, after which the scores stay relatively constant. Therefore, a value of 9 is decided, which is significantly higher than the value of 1 used by Shi et al. (2018), Marinov et al. (2020), and Chichirau et al. (2020). The tuning process of α is in line with the process of Shi et al. (2018), where the value was also increased until no significant benefit was observed. It should be noted, however, that as stated by Quan et al. (2015), a better topic coherence (PMI) does not guarantee overall better quality.

3.3 Running the model

Running the model with $n_t = 100$, $K = 50$, and $\alpha = 9$ results in a mean Topic Score and mean Pointwise Mutual Information score of 0.996 and 2.80 respectively. Training took four and a half hours while using over 55GB of memory. The PMI score indicates a high level of homogeneity within topics. Secondly, the TD indicates a high level of heterogeneity between topics. The PMI score is comparative to the PMI score obtained in other work regarding tweet clustering using the SeaNMF model (Marinov et al., 2020; Chichirau et al., 2020; Shi et al., 2018). Marinov et al. (2020) indicates that generally the PMI score tends to decrease with the number of documents. Because of the relatively large number of documents used in this research, a PMI score of 2.80 indicates the resultant topics likely represent specific subtopics of COVID-19.

3.4 Output Formatting

Running the SeaNMF algorithm with the parameters described in Section 3.2.3 provides us with 50 individual topics. For proper analysis of topic

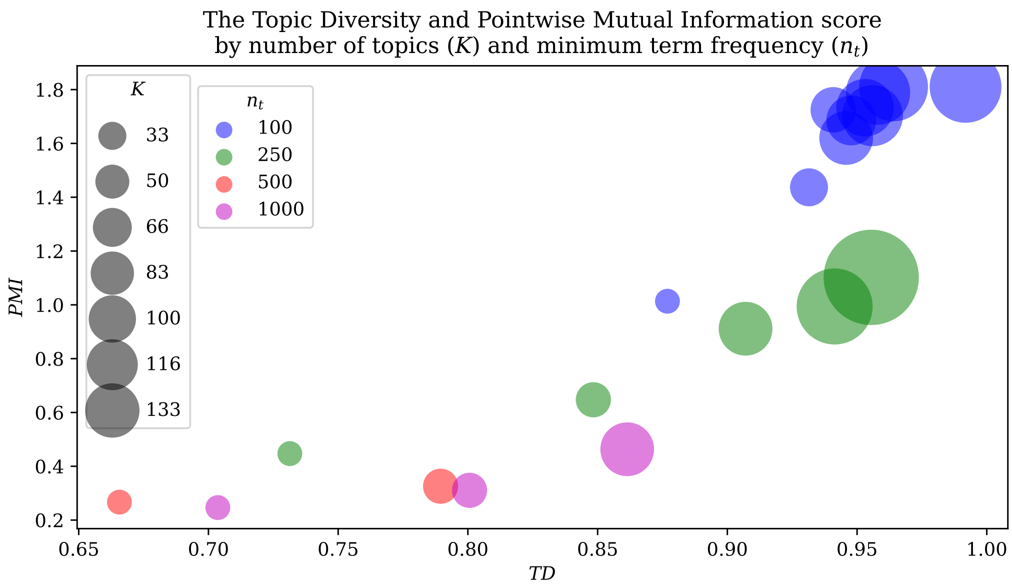

Figure 3.1: The results of the tuning process of the SeaNMF model for different values of n_t and K with $\alpha = 1$. The TD indicates topic uniqueness and the PMI score indicates the level of topic coherence. A lower n_t and a larger K together result in higher TD and PMI scores.

differences, topics are combined into topic categories. This categorization is important because between some topics similarity is observed by us that was not picked up by the model. Furthermore, decreasing the number of topics makes analysis more intuitive. Because the document-level word co-occurrence was not accurately captured for topics within a category, categorization is done by hand. For the same reason, the PMI score and TD cannot be used for judging category soundness.

Topics with a PMI lower than 0.5 are not assigned to any category, because a low PMI indicates the topic does not effectively capture a subtopic. Furthermore, some topics with a high PMI are removed, because they are judged to be irrelevant. For example, a topic containing only English stop words is removed. A topic is not exclusively assigned to a single category as well. For example, a topic about governmental financial support belongs to both measures & politics and economy & business.

Matrix H (Equation 2.4) contains the correlations between all topics and tweets. Using H, we can derive a tweet-category matrix C as the resultant of all score vectors from the categories assigned to that vector (Equation 3.4).

$$\vec{s}_{category} = \sum_{i=1}^{n} \frac{\vec{s}_i}{n} \tag{3.4}$$

In Equation 3.4 the score vector for every category $\vec{s}_{category}$ is normalized for the number of topics n in each category. Finally, each day is assigned a day score for every category as the mean of the total topic scores for all tweets of that day.

4 Results

To properly investigate the effect of public addresses on public discourse on Twitter, it is important to make a distinction between the two goals of public addresses. Firstly, the press conferences served both as a way for the government to reveal the measures decided by the government. This will be referred to as the objective goal. A second goal of the press conference was to maintain support for the measures, make sure people feel the importance of complying with the measures and also consolidation for those feeling anxious (van Weelden, 2020). This will be referred to as the subjective goal. Hoogstraten (2020) explains the obedience of the measures by people not part of the high-risk group by a feeling of altruism, illustrating the importance of the subjective goal. The three public addresses that were not press conferences, the two speeches at the Torentje and the kings speech, can be regarded as purely subjective, as no new measures were announced.

Although the data provided by the model doesn't differentiate between these goals, the interpretation of the data does depend on which goal is considered. Regarding the effect of the objective goal, most measures weren't initially publicized during the press conferences, but rather by media outlets during the days before (Mulder, 2020). Because of this, a change in relative number of tweets a few days before a press conference can still be allocated to the objective goal of that press conference. Regarding the subjective goal, the change would be expected on the day of or the days after a press conference.

The standard score of a day over a window of the 10 previous days ($Z_{\delta x=10}$) indicates the size of the change in day score for a day compared to the 10 days before. For every day, the standard score is determined using a window of the 10 days up to that day. A windowed standard score of 0 then means no change was observed and a windowed standard score of 1 means the day score is 1 standard deviation higher than the mean day score over that window. For every day, a day score is assigned per category, representing the relative number of tweets about that category. The day score variability is generally higher during the initial months of the pandemic (Section 4.1). The level of change that correlates with public addresses is limited. In general, changes in topic category discussion around the days of public addresses can mostly be regarded to a more widespread trend observed (Section 4.2).

4.1 General Trends

A high level of fluctuation in topics discussed between February and December is observed. This is illustrated by the relative standard deviation (RSD) of all separate categories in Table 4.1. The RSD represents the size of the standard deviation compared to the mean. The travel and leisure category has the highest RSD value, which means that the change in the amount of discussion of this category on Twitter was the highest. The RSD score for the topic of measures and politics is relatively small, meaning that the percentage of tweets regarding this category stays relatively similar over the course of the year.

Category Title	RSD	$\overline{RSD}_{\delta x=10}$
Research	22.4%	11.1%
Cases & Healthcare	22.3%	11.8%
Measures & Politics	13.2%	8.1%
Economy & Business	19.1%	15.3%
Culture & Sport	19.2%	13.4%
Travel & Leisure	31.4%	16.6%

Table 4.1: The relative standard deviation of all day scores and the mean relative standard deviation over a 10 day window per category.

Furthermore, the mean of all relative standard deviations over a window of 10 days indicates a high level of fluctuation for shorter durations as well ($\overline{RSD}_{\delta x=10}$, Table 4.1). The mean windowed RSD for the topic regarding economy and business is relatively high compared to its total RSD. This means there was often a high change in the percentage of tweets about the economy and business during a given ten-day period. The windowed RSD for the category regarding measures and politics is the smallest of all categories, meaning that not only the change in tweets about the category over the year was relatively small, but also during a shorter time span.

In Figure 4.1, the mean of all day scores is presented. Because of the high level of variation, the data is presented over a 10 day running mean to increase readability. The remainder of this subsection describes the general trends observed per category.

Research – For most days, the category of research is the most discussed. Three peaks are observed for the category. The first coincides with the first officially identified cases in The Netherlands (National Institute for Public Health and the Environment, 2020). The second peak occurs around the time that worries arise about the possibility of new variants of the virus, partly originating from mink farms (Nederlandse Omroep Stichting, 2020). The third peak coincides with the second national lockdown and the initial press releases about the probable sufficient efficacy of different vaccines (Erman and Steenhuysen, 2020). The minister of health has

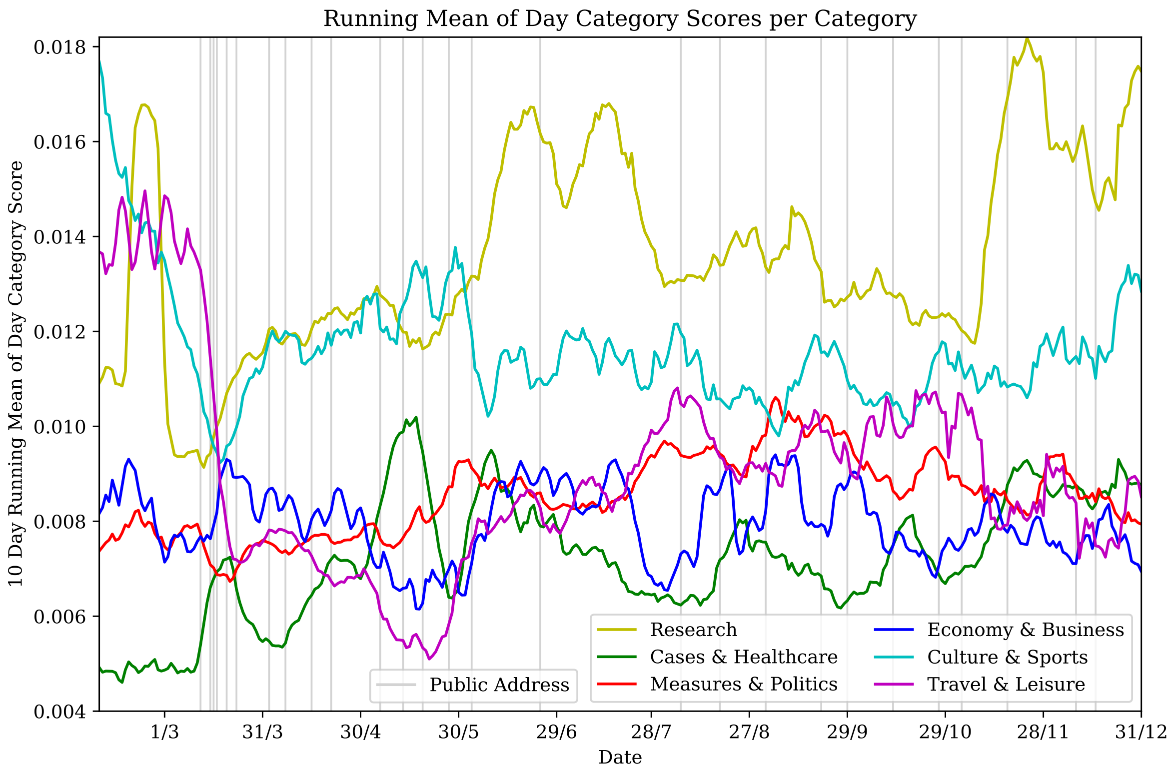

Figure 4.1: The 10 day running mean of the day score per topic category for all tweets of each day indicates the activity of the category discussion. A higher score represents a higher average correlation between all tweets for a day and the topics that are part of the category score.

pronounced vaccinations to be the only way out of the pandemic on multiple occasions, including press conferences (College ter Beoordeling van Geneesmiddelen, 2020; Ministerie van Algemene Zaken, 2020b). Vaccine safety was also highly covered by both traditional news outlets and fake news websites, causing great societal discussion about the topic (Steenbreker, 2020).

Culture & Sports – Initially, there is a high number of tweets regarding culture and sports. This number decreases during the months of February and April, increasing quickly again afterward. This increase coincides with the announcement of the cancellation of all events for over 100 people. Adaptions on international sporting events during February and March can explain the generally high level of discussion for these categories during this period (Ministerie van Algemene Zaken, 2020a).

Cases & Healthcare – The initial number of tweets about cases and healthcare is relatively low. The awareness of the potentially high number of COVID-19 cases in The Netherlands during the month of February and the first half of March was low, which correlates with the low number of tweets about cases and healthcare during this period. An increase in day score is observed during the rise in the national infection rate. During the summer months, when infections were low, we notice a small general decrease again (Rijksoverheid, 2021).

Measures & Politics – The number of Tweets concerning Measures and Politics has the least variability both over the course of the year as on shorter time spans. The percentage of tweets about measures & politics is observed to stay relatively similar every day compared to the other topics, which can also be observed from the relatively low RSD and windowed RSD.

Travel & Leisure – As international travel was initially obstructed, the number of tweets about travel is observed to be at its highest. Later, during the summer months, as restrictions on travel were loosened, an increase in Tweets about travel is observed as well (Ministerie van Algemene Zaken, 2020c).

Economy & Business – The number of tweets about the economy and business stays relatively equal over the course of the year. Compared to the topic's RSD value, the windowed RSD is relatively large. This indicates a high fluctuation in tweets about this category per day.

4.2 Individual Public Addresses

The standard scores for the five days before and the five days after each press conference are displayed in Figure 4.2. A stark increase or decrease in standard score on the days around a public address indicate the public address might have influenced the number of tweets about that category. Figure 4.1 specifies whether the change in standard score on a day reflects an increase or decrease of tweets about a category.

For most days surrounding public addresses, we notice no or a negligible change in standard score, indicating there is no general effect of public addresses on a single topic. For some public addresses, however, a large change in standard score for a specific topic is observed. The public addresses that correlate with substantial difference in change are described below[‡].

March 12th - March 23rd – During this period, press conferences are held frequently. On March 12th, the first official COVID-19 specific press conference was held. This included the announcement of an intelligent lockdown, which entails the start of working from home and the cancellation of all events for over 100 people. An increase in tweets regarding measures and politics is observed on the day of and days after the press conference. On the days after the press conference, there is an increase in tweets about research as well. Lastly, there is a decrease in tweets regarding travel and leisure. Figure 4.1 indicates this decrease is part of a larger trend, however. Therefore, it is unlikely that the press conference affected the number of tweets about this category significantly. A few days after the first press conference, a new press conference is held. Here, the closing of schools, bars & restaurants, and gyms is announced and current measures are extended until April 6th. After this press conference, the variability in standard scores decreases for all topics. Notably, the change in standard score stays low after the public address by the prime minister at the Torentje on March 16th. On the 17th of March, Financial support for companies and the self-employed is announced. Notably, this does not correlate with a change in number of tweets regarding the economy and business.

May 27th – At this press conference, people are told to be cautious with going on vacation abroad and gyms, casino's and saunas are allowed to open. Furthermore, organized sports is now allowed outside for people under the age of 19. On the days before the press conference, the change in standard score for the categories of culture and sports, and research is minimal. On the days after the press conference, there is a sudden drop in tweets about culture and sports.

September 18th – On the day of this press conferene, a small increase in discussion about measures and politics is observed. Increases of this level, however, are observed to be frequent for this category (Figure 4.1).

October 13th – The second lockdown is announced: face masks are obligatory in public spaces, alcohol is banned after 8 p.m., and the number of people allowed in stores is limited. This coincides with a short lived increase in discussion regarding measures and politics.

December 14th – During this press conference, the measures regarding the second lockdown are tightend, limiting the ways people can come together during Christmas. On the day of and the days after, a stark increase in discussion regarding travel and leisure is observed.

5 Discussion

The aim of this study was to investigate whether the SeaNMF model can be used to analyze how governmental public addresses during the COVID-19 pandemic might have influenced public discourse on Dutch Twitter. The PMI and TD metric indicate a sufficient model quality. Over the course of the year,

[‡]Information regarding the content of public addresses is provided by Ministerie van Algemene Zaken (2021)

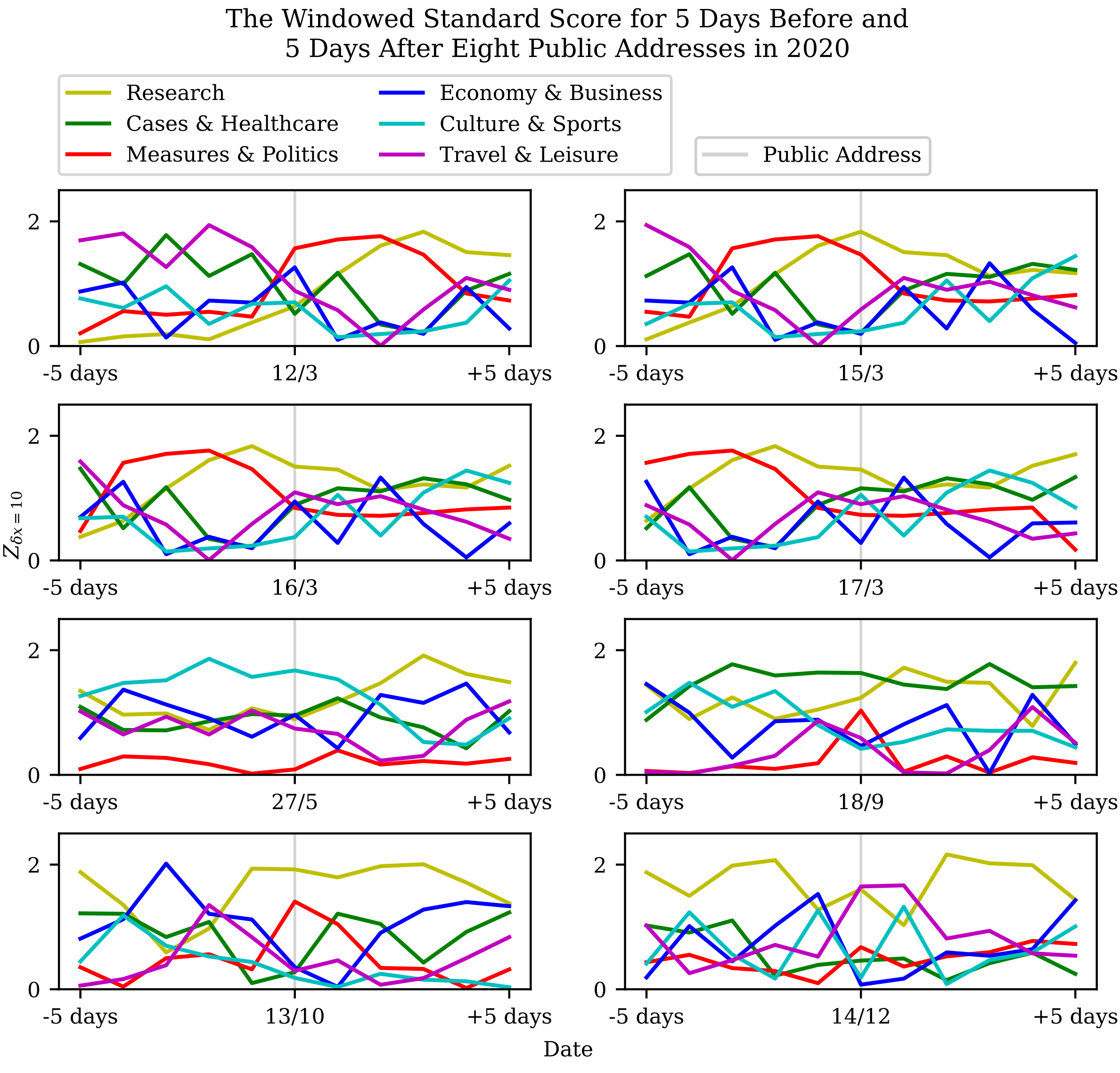

Figure 4.2: The change in day score by topic per day compared to the previous 10 days represented as the windowed standard score. The graphs for all public addresses of 2020 are presented in Appendix B.

changes in topic discussion that coincide with real world events are observed, indicating valid usage of the SeaNMF model for modelling long term trends (Section 5.1). The observed relation between public addresses and topics discussed is however minimal, indicating the SeaNMF model is limited for modelling short term changes over large datasets (Section 5.2).

5.1 Changes in Long Term Topic Discussion

The extremes in daily topic discussion are observed to be higher for the months of February and March compared to the rest of the year. The number of tweets per month are equalized during preprocessing, ensuring this difference is not caused by a relatively low amount of data. It is possible the initially high level of fluctuation in topic discussion is caused by the general feeling of uncertainty regarding the pandemic during this period.

The general relative level of category discussion differs during these months as well. The latter is likely the result of public discussion changing from a focus on global issues to national issues, as observed by Marinov et al. (2020).

Most general changes in topic discussion over the course of the year can be attributed to long term events regarding the pandemic, indicating a sufficient quality of the model for the purpose of investigating long term trends. Furthermore, metrical evaluation of the topics returned by the model using the TD and PMI score indicate an overall good quality of the model.

5.2 The Effect of a Public Address on Topic Discussion

A generally high change in category discussion is observed for most days, making it difficult to discern the differences in day scores for days surrounding public addresses. Contrary to these general findings, however, the first few public addresses do correlate with changes in topic discussion for some categories. This indicates the possibility of public addresses having significant influence on category discussion during the initial months of the pandemic. It should be noted, however, that during this period a generally higher level of variability per category is observed as well. Because of this, it is likely the correlation between public addresses and topic discussion during this period is the result of chance. Overall, the results are inconclusive.

Some public addresses do correlate with stark changes in topic discussion. It should be noted however, that these correlations are generally sparse. For most public addresses, no significant effect on public discourse on Twitter is observed. This indicates the SeaNMF model is limited for analyzing trends in small parts of a large dataset. In this use case, that is daily trends over the course of the year. Three possible reasons for this are explained below.

5.3 Model validity

Firstly, in order to decrease the computational load of the model, about 92% of documents were removed from the dataset, resulting in an average of less than 1500 tweets per day available for topic scoring and categorization. Although the total number of documents provided a substantial amount of data, the number of daily documents might have been too low to accurately group the level of discussion per category on a daily basis. The possibility of this can be observed from the high level of variability in day score. It is not clear whether the high variability in shorter term day scores is inherent to the data, or caused by data sparsity. To investigate this, one could run the model over a shorter time span on the same data set, but without stripping the data. If the observed variability is significantly lower in this test, it can be concluded that the variability here is indeed caused by data sparsity. Possibilities to limit data sparsity are discussed in Section 5.4.

Another cause for the absence of a relation can be the quality of the topic categories. During the topic categorization process, many assumptions were introduced. It is therefore highly likely that categories either encapsulated a too broad or too narrow range of topics. The correlation between the number of tweets regarding COVID-19 related research and events in COVID-19 related research indicates this category accurately captures the subtopics. Regarding other categories, the same level of correlation is not observed by us. If those categories accurately capture their subtopics, this simply indicates that the change in discussion over the year was not impacted by their related events. However, the small level of correlations also indicates the categories were too broadly defined, therefore not accurately presenting the level of change. This problem can be best handled by eliminating the need for topic categorization all together (5.4).

The inconclusive results are not necessarily the result of the possible limits of the SeaNMF model for our use case. As observed by Chichirau et al. (2020), many different events might influence topic discussion during the pandemic. This could mean that the public addresses did influence public discussion to some extent, however the number of other events that influence the level of short term topic discussion is so high, that distinguishing between the effect of public addresses and other events is impossible given our data. To investigate the possible effect of public addresses on public discussion further, it can therefore be helpful to take other events into account as well. To properly distinguish between multiple events, it may be necessary to shorten the time span used for document grouping, which was a day in this research. Decreasing this time span, however, will likely increase the problem of data sparsity as the available number of documents per time span is consequently reduced.

5.4 Future Research

Regarding future research, this subsection goes over possible different topic modelling techniques in order to decrease data sparsity and bias. Afterwards, researching other parts of the relation between public discourse and public addresses not encapsulated by our methodology are discussed.

The computational load required by the model resulted in its limitations for larger corpora. Since these limitations are inherent to the design of the NMF and the SeaNMF model, other models might be better suited for the large dataset used in this research. In recent years, significant advances have been made in the quality of publicly available word vector embeddings and other language representation models for the use of topic modelling (Devlin et al., 2018; Zhao et al., 2021; Peinelt et al., 2020). The BERT model introduced by Devlin et al. (2018) has been successful in topic discovery and sentiment classification on social media posts regarding COVID-19 (Pandey, 2021). These results indicate that using a different topic modelling technique can significantly improve the output of the model.

The Tiny Belief Propagation (TBA) model introduced by Zeng et al. (2012) successfully reduces the memory complexity of topic modelling. This technique has not been properly tested on short documents, however, and since TBA utilizes document-level word co-occurrence as well, we expect the problem of a lack of semantic patterns available to be apparent here too. Using an external resource to incorporate semantic information into the model has been shown to increase the models performance for both topic modelling based on NMF and LDA (Nguyen et al., 2015; Le and Mikolov, 2014). Because of this, the TBA model might return satisfactory results when external additional semantic information is introduced.

The SeaNMF model is also limited by the bias introduced during categorization, pre-processing, and pre-defining the number of topics K. While the PMI and TD scores indicate which number of topics return the best results, the model itself does not discover the number of topics inherently present in the data. Niu et al. (2015) proposes a topic modelling technique based on clustering the word embeddings into a multidimensional vector space. Documents and terms are embedded using an external library, for example the BERT Sentence Transformer introduced by Devlin et al. (2018). Consequently, the vectors are mapped into lower dimensional embedding. This effectively means that semantically similar terms and documents are grouped close together in the vector space. Documents and terms that correlate with a wide range of other terms and documents, for example stop words, are not closely positioned to any cluster centroid, making them easily distinguishable outliers. As a consequence, the number of topics is discovered automatically and pre-processing steps like lemmatization and stop-word removal, are not necessary. Furthermore, the method has been proven effective on short texts (Niu et al., 2015). Using this model can therefore increase the integrity of the evaluated topics.

Lastly, this research focused on the effect of the public addresses on public discourse, however the effect of public discourse on the public addresses can also provide relevant insights in the relation between the two. Many pieces are written about the cabinet following public discourse in designing its COVID-19 regulations to a large extent (Pré, 2020; Valk and Rusman, 2020; Vrieze, 2020; Vosselman and De Loo, 2020). Moreover, some relation between public opinion and the public addresses' message would not be out of place for a democratically elected government. Because of this, a bidirectional relation between public addresses and public discourse might have influenced our results. For example, if the content of a public address is partly related to prior public discourse, this might decrease the correlation between day score change and public addresses. A proper investigation to this effect can provide further insights in the relation between the two.

5.5 Conclusion

In this paper, the relation between governmental public addresses regarding COVID-19 and topic discussion on Dutch tweets is investigated using the SeaNMF model. General trends in topic discussion over the year are observed. A large difference in general topic discussion between February and March, and the rest of the year is observed. The number of tweets regarding COVID-19 related research is generally the highest over the course of 2020, showing significant increases during the summer months and the end of the year. For other topic categories, the change in number of tweets over the course of the year is relatively small.

There is a high level of daily variability in topic discussion. Our results indicate some public addresses could have influenced topic discussion to some extent, however the level of variability in the data is generally too high to draw conclusive results. Therefore, while no general effect of public addresses on public discourse is observed, no conclusion about the lack of this effect can be made. We propose including semantic information from an external source, for example BERT, and clustering the documents using a transformer based model.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Buschow, C., Schneider, B., and Ueberheide, S. (2014). Tweeting television: Exploring communication activities on Twitter while watching TV. *Communications*, 39(2).

Caselli, T. and Basile, V. (2020). 40twene_nl. `https://osf.io/pfnur/`.

Chichirau, M., Spenader, J., and Caselli, T. (2020). Topic modelling and emotion detection on Italian tweets during the early Covid-19 pandemic.

Cichocki, A., Zdunek, R., Phan, A. H., and Amari, S.-i. (2009). *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation.* John Wiley & Sons.

College ter Beoordeling van Geneesmiddelen (2020). Het grootste coronaspreekuur - medicijnen & vaccins.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Erman, M. and Steenhuysen, J. (2020). Pfizer vaccine trial success signals breakthrough in pandemic battle.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*. ACM Press.

Hoogstraten, R. v. (2020). Veiligheid boven vrijheid? een interdisciplinair onderzoek naar verklaringen voor de publieke naleving van coronamaatregelen in Nederland. Master's thesis, Utrecht University.

Jungherr, A., Schoen, H., Posegga, O., and Jürgens, P. (2016). Digital trace data in the study of public opinion. *Social Science Computer Review*, 35(3):336–356.

Kuang, D., Ding, C., and Park, H. (2012). Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China. PMLR.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Lin, X. and Boutros, P. C. (2020). Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics*, 21(1).

Marinov, B., Spenader, J., and Caselli, T. (2020). Topic and emotion development among Dutch COVID-19 Twitter communities in the early pandemic.

Ministerie van Algemene Zaken (2020a). Letterlijke tekst persconferentie minister-president Rutte en minister Bruins naar aanleiding van de maatregelen tegen verspreiding coronavirus in Nederland.

Ministerie van Algemene Zaken (2020b). Letterlijke tekst persconferentie minister-president Rutte en minister De Jonge (1-9-2020).

Ministerie van Algemene Zaken (2020c). Persconferentie coronavirus: zomervakantie 2020 in eenvoudige taal.

Ministerie van Algemene Zaken (2021). Video's en letterlijke teksten persconferenties coronavirus.

Mulder, J. (2020). Xander van der wulp over wie de maatregelen lekt voor de persconferenties van rutte.

National Institute for Public Health and the Environment (2020). Patiënt met nieuw coronavirus in Nederland.

Nederlandse Omroep Stichting (2020). Opnieuw twee nertsenfokkerijen besmet met coronavirus in brabant.

Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Niu, L., Dai, X., Zhang, J., and Chen, J. (2015). Topic2Vec: learning distributed representations of topics. In *2015 International conference on asian language processing (IALP)*, pages 193–196. IEEE.

Pandey, C. (2021). redBERT: A topic discovery and deep sentiment classification model on COVID-19 online discussions using BERT NLP model.

Peinelt, N., Nguyen, D., and Liakata, M. (2020). tBERT: Topic models and BERT joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055, Online. Association for Computational Linguistics.

Pré, R. d. (2020). Weer buigt Rutte blijmoedig voor de publieke opinie. maar van een weldoordacht plan is nog geen sprake.

Quan, X., Kit, C., Ge, Y., and Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 2270–2276. AAAI Press.

Rijksoverheid (2021). Positief geteste mensen: Coronadashboard: Rijksoverheid.nl.

Shi, T., Kang, K., Choo, J., and Reddy, C. K. (2018). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. ACM Press.

Sridhar, V. K. R. (2015). Unsupervised topic modeling for short texts using distributed representations of words. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Association for Computational Linguistics.

Steenbreker, B. J. (2020). Advocating hysteria : How Dutch media used framing and conspiracy coverage during the corona virus of 2020 compared to the swine flu of 2009.

Stichting Kijkonderzoek (2021). Kijken in tijden van corona - Kerncijfers tv 2020. `https://kijkonderzoek.nl/images/Persberichten/2020/Jaarpersberichten/210107_Factsheet_coronacrisis_-_kerncijfers_TV_2020.pdf`. Accessed: 22-5-2021.

Valk, G. and Rusman, F. (2020). Kabinet laat de volkswil meetellen in de coronacrisis.

van Weelden, E. (2020). Een kwalitatief onderzoek naar de rol van crisiscommunicatie door de overheid tijdens persconferenties over Covid-19 in het preventiegedrag van Nederlandse jongeren. Master's thesis, Utrecht University.

Vosselman, E. and De Loo, I. (2020). Performativiteit in wetenschap en accountancy. *Maandblad Voor Accountancy en Bedrijfseconomie*, 94:181.

Vrieze, J. d. (2020). Waarom het vertrouwen in het coronabeleid afbrokkelde.

Xun, G., Gopalakrishnan, V., Ma, F., Li, Y., Gao, J., and Zhang, A. (2016). Topic discovery for short texts using word embeddings. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE.

Zeng, J., Liu, Z.-Q., and Cao, X.-Q. (2012). Memory-efficient topic modeling.

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., and Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing Twitter and traditional media using topic models. In *Lecture Notes in Computer Science*, pages 338–349. Springer Berlin Heidelberg.

A Public Addresses and their Content

Date	Description
12/3	Announcement of the intelligent lockdown: work at home and cancellation of events for over 100 people.
15/3	The losing of schools, bars & restaurants, and gyms is announced. Current measures are extended until April 6th.
16/3	Public address by the prime minister at the Torentje. Achieving herd immunity is mentioned as strategy.
17/3	Financial support for companies and the self-employed is announced. UEFO EURO 2020 is postponed til next year.
20/3	The speech by the Dutch king.
23/3	Existing measures are tightend. All events are cancelled until June first.
31/3	Extension of all measures till at least April 28th.
7/4	The capacity of COVID-19 tests is increased. There are talks about the development of the COVID-19 app.
19/5	The measures affecting the hospitality industry are relaxed.
15/4	Fincancial support for the cultural sector is announced. The one-and-a-half-meter econonmy is discussed.
21/4	The future opening of schools is announced, otherwise the lockdown is extended. No events are allowed until September first.
6/5	Announcement of the initial relaxation of measures. Face masks are introduced in public transport.
13/5	Rapid COVID-19 tests are critized. A general feeling of ambiguity about the summer vacation is expressed by the journalists.

Date	**Description**
27/5	People are told to be cautious with going on vacation abroad. Gyms, casino's and saunas are allowed top open.
3/6	Regulations regarding traveling abroad are elaborated.
24/6	Public transport allowed for non-essential journeys.
6/8	The Dutch Broadcasting Foundation (NOS) analyzes this press conference to be a warning to the public.
18/8	First new measures for the second lockdown: no more than 6 guests are allowed per house.
1/9	The safety regions are announced to be playing a larger role in containing the spread of COVID-19.
18/9	Regional measures are imposed.
28/9	Working from home becomes the norm again, public is not allowed at sport events and bars close at 10 p.m..
13/10	The second lockdown is announced: face masks are obligatory in public spaces, alcohol is banned after 8 p.m., the number of people allowed in stores is limited.
27/10	Current measures are extenden until December.
3/11	Cinema's, museums, and zoo's are closed. The possibility of a local curfew is mentioned.
17/11	A slight relaxation of measures is announced. The beginning of 2021 is meentioned as a possible start of vaccination.
8/12	Januari is indicated as a month were more relaxation of measures could be announced. Vaccination is mentioned as the only solution.
14/12	Public address by the prime minister at the Torentje: a hard lockdown is announced.

B Public Addresses and their Windowed Standard Score

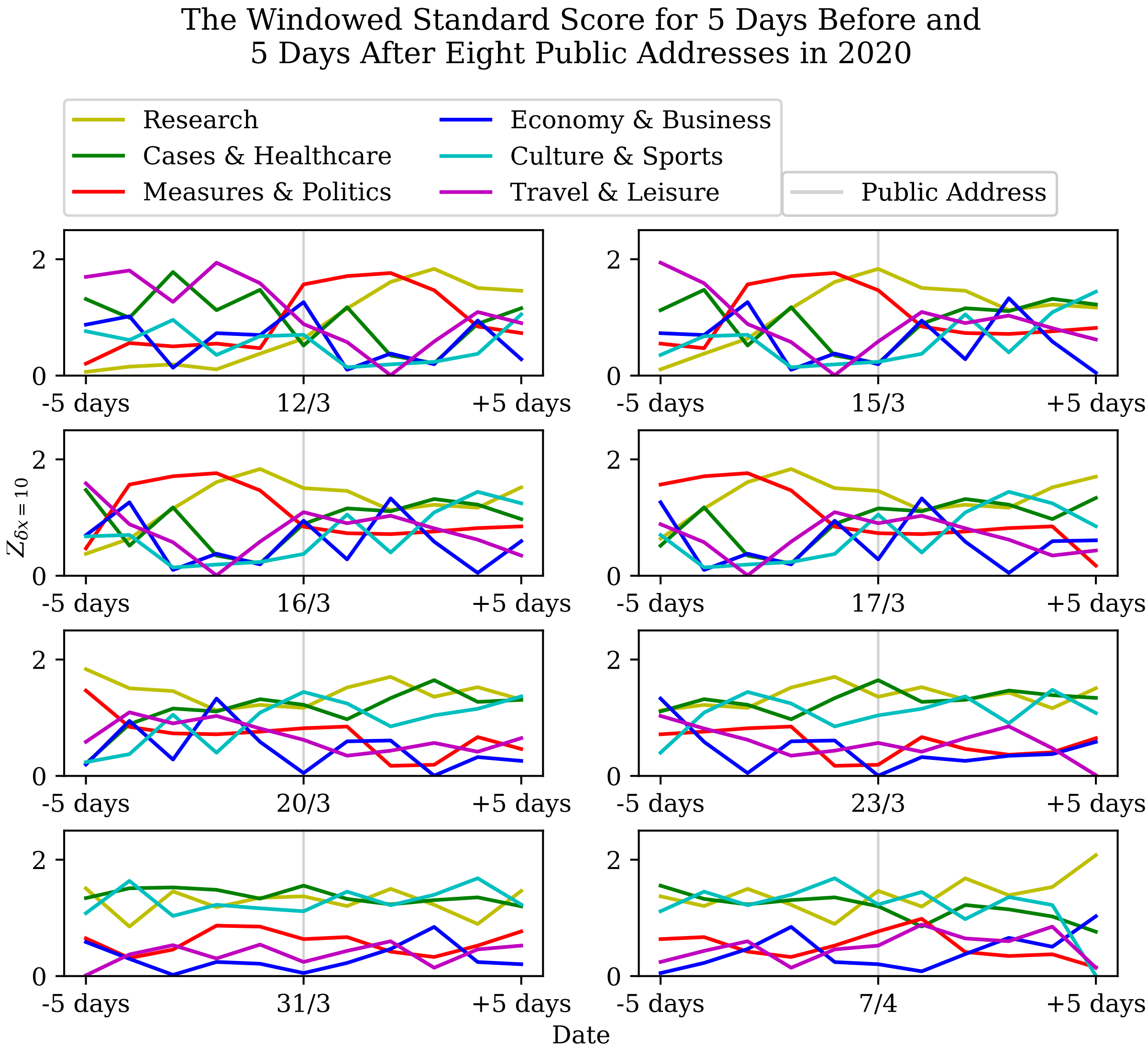

Figure B.1: The change in day score by topic per day compared to the previous 10 days represented as the windowed standard score.

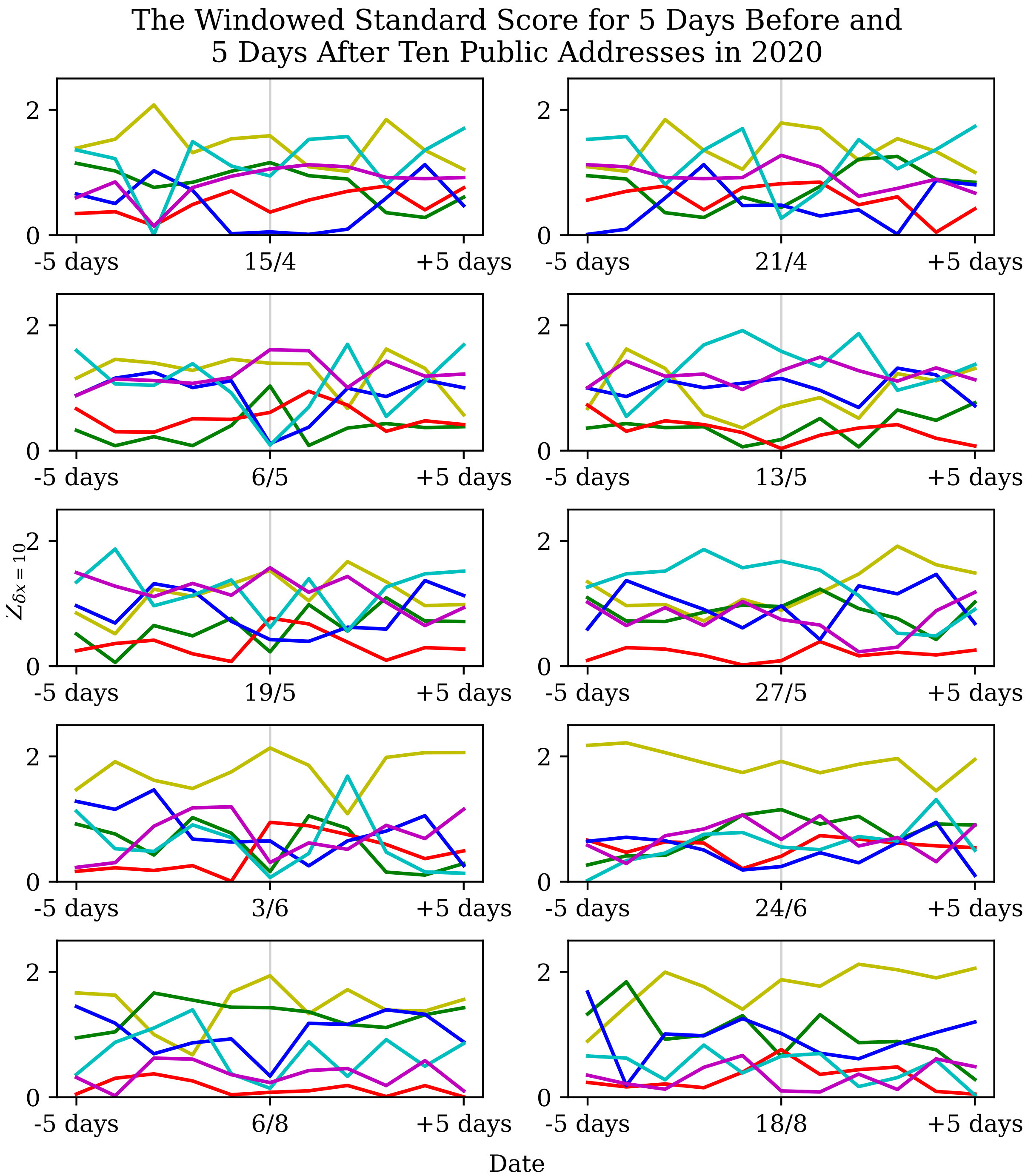

Figure B.2: The change in day score by topic per day compared to the previous 10 days represented as the windowed standard score. The legend is presented in Figure B.1.

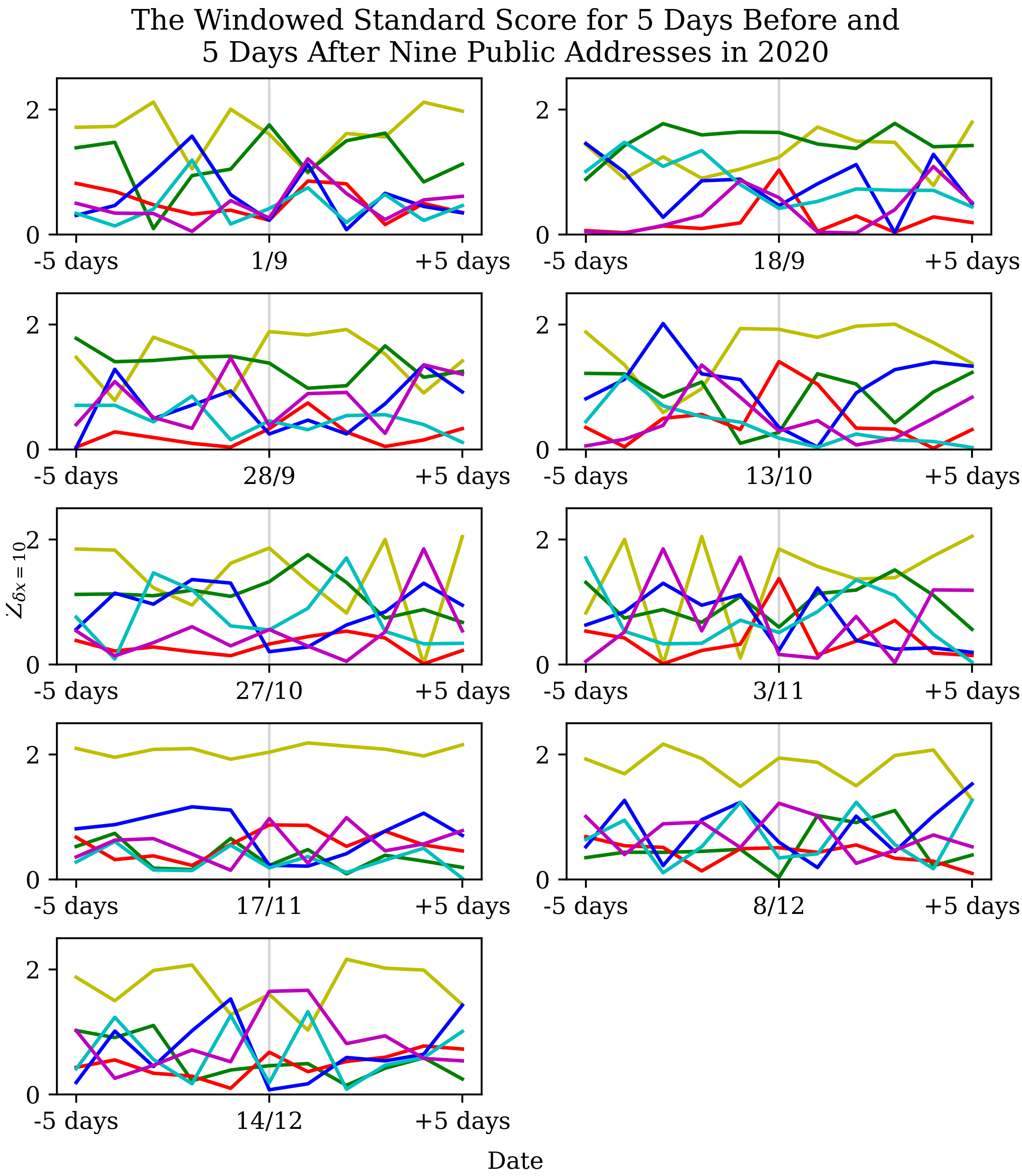

Figure B.3: The change in day score by topic per day compared to the previous 10 days represented as the windowed standard score. The legend is presented in Figure B.1.