Topic modelling of Finnish Internet discussion forums as a tool for trend identification and marketing applications

Ilkka Särkiö

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.
Helsinki 13.2.2019

Supervisor

Assistant Prof. Pauliina Ilmonen

Advisor

M.Sc. (Tech) Mikko Koski

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Ilkka Särkiö

Title Topic modelling of Finnish Internet discussion forums as a tool for trend identification and marketing applications

Degree programme Mathematics and Operations Research

Major Systems and Operations Research **Code of major** SCI3055

Supervisor Assistant Prof. Pauliina Ilmonen

Advisor M.Sc. (Tech) Mikko Koski

Date 13.2.2019 **Number of pages** 78+35 **Language** English

Abstract

The increasing availability of public discussion text data on the Internet motivates to study methods to identify current themes and trends. Being able to extract and summarize relevant information from public data in real time gives rise to competitive advantage and applications in the marketing actions of a company. This thesis presents a method of topic modelling and trend identification to extract information from Finnish Internet discussion forums. The development of text analytics, and especially topic modelling techniques, is reviewed and suitable methods are identified from the literature. The Latent Dirichlet Allocation topic model and the Dynamic Topic Model are applied in finding underlying topics from the Internet discussion forum data. The discussion data collection with web scarping and text data preprocessing methods are presented. Trends are identified with a method derived from outlier detection. Real world events, such as the news about Finnish army vegetarian meal day and the Helsinki summit of presidents Trump and Putin, were identified in an unsupervised manner. Applications for marketing are considered, e.g. automatic search engine advert keyword generation and website content recommendation. Future prospects for further improving the developed topical trend identification method are proposed. This includes the use of more complex topic models, extensive framework for tuning trend identification parameters and studying the use of more domain specific text data sources such as blogs, social media feeds or customer feedback.

Keywords Topic modelling , Social Media , Natural Language Processing , Text Analytics , Text Mining , Trend identification , Digital marketing

Aalto University
School of Science

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Diplomityön tiivistelmä

Tekijä Ilkka Särkiö

Työn nimi Suomalaisten internetkeskustelufoorumien aihemallinnus trendintunnistuksen ja markkinoinnin sovellusten työkaluna

Koulutusohjelma Matematiikka ja operaatiotutkimus

Pääaine Systeemi- ja operaatiotutkimus **Pääaineen koodi** SCI3055

Työn valvoja Assistant Prof. Pauliina Ilmonen

Työn ohjaaja M.Sc. (Tech) Mikko Koski

Päivämäärä 13.2.2019 **Sivumäärä** 78+35 **Kieli** Englanti

Tiivistelmä

Internetissä on enenevissä määrin saatavilla julkista tekstimuotoista dataa, mikä motivoi tutkimaan menetelmiä, joilla tekstistä voi tunnistaa ajankohtaisia aiheita ja teemoja. Kyky erottaa ja tiivistää merkityksellistä tietoa julkisesta datasta reaaliajassa antaa yrityksille kilpailuedun ja mahdollisuuksia sovelluksiin markkinoinnissa. Tässä opinnäytetyössä esitellään aihemallinnukseen ja trendintunnistukseen perustuva menetelmä tiedon löytämiseksi suomalaisista internetkeskustelufoorumeista. Työssä tarkastellaan tekstianalytiikan ja erityisesti aihemallinnuksen kehityksen kirjallisuutta sekä tunnistetaan soveltuvia menetelmiä. Latent Dirichlet Allocation -aihemallia sekä dynaamista aihemallia (Dynamic Topic Model) sovelletaan työssä keskustelufoorumeiden tekstin latenttien aiheiden tunnistamiseen. Tekstidatan keräämistä hakurobotin avulla sekä tekstidatan esikäsittelyä kuvataan työssä. Tunnistettujen teemojen trendejä ja erityisiä ajankohtia tunnistetaan työssä kehitetyllä menetelmällä, joka perustuu poikkeavien havaintojen tunnistustekniikoihin. Aiheista tunnistettiin ilman ulkoista ohjausta oikeita, ajankohtaisia tapahtumia kuten Suomen puolustusvoimien kasvisruokapäivän ehdottaminen sekä presidentti Putinin ja Trumpin huippukokous Helsingissä. Työssä tarkastellaan myös löydetyn tiedon soveltamista markkinointiin esimerkiksi hakukonemainonnan automaatioon sekä sisältöjen suositteluun liittyen. Trendintunnistuksen menetelmän jatkokehityksen suuntia tarkastellaan esimerkiksi monimutkaisempien aihemallien sekä järjestelmällisen parametrivalinnan kannalta. Myös mahdollisten muiden datalähteiden käyttöä arvioidaan.

Avainsanat Aihemallinnus, sosiaalinen media, luonnollisen kielen analyysi, tekstianalytiikka, trendien tunnistus, digitaalinen markkinointi

Preface

As I began my studies in the program of Technical Physics and Mathematics in Aalto University in September 2011, I could not have believed to be completing my Master's thesis on a subject that is most related to Computer Science and moreover, in the domain of marketing. As of now, I really want to tell all aspiring young students about the vast possibilities that lie in the STEM field. I feel confident that everything I have experienced and learned will guide me to an interesting career filled with possibilities. I thank my family in their support and persistent interest in my studies and student life. You have been a necessary part of my success.

My professor, academic advisor and thesis supervisor Pauliina can not be given enough appraisal for her incredible ability to induce enthusiasm and interest in mathematics, studying and for me, this thesis. I thank you for your support, amazing courses and our regular thesis meetings.

I thank Tiina for first introducing me to the Advanced Analytics team at Dagmar. Working here and completing my thesis has been an irreplaceable experience that has aided me a lot in getting the most out of my Master's studies. I thank my boss Mikko for giving me inspiration and support for the lengthy process of writing this thesis. Special thanks goes to Teemu for providing the data necessary for this thesis.

I address my sincerest gratitude to the amazing community at Aalto University in supporting my learning, both in studies and about life, throughout the years of my study. I grew to believe in my talents and gained confidence to pursue learning in many different fields. The Guild of Physics and my dear friends there have especially been an inspiration and an important resource to carry on studying, despite difficult courses and periods of lost motivation. My friends at Raati3 and Raati15 have become lifelong friends. Volunteering at the Aalto University Student Union has given me invaluable skills in project management, experience in managing an organization and especially friends and contacts for my future life. The board of 2016 and the year we spent together will remain dear to me for years to come.

Besides important life lessons and career relevant skills I obtained during my studies and time as in the Aalto community, the fun I have had during my years in Otaniemi is something to treasure for the rest of my life. The most bizarre and memorable student life moments in Otaniemi I have to credit to Vapaateekkarit. Being a part of this amazing bunch of people has been and will continue to be an inseparable part of me. At this time of graduating from my school, I am sure that I truly have experienced the best student life possible.

Lastly, I thank my dear partner Anna in giving me daily confidence, support and love. Without our long conversations on the couch, in between cooking our meals or during car rides to wherever, I doubt I could have achieved what I currently have.

Helsinki, 13.2.2019

Ilkka Särkiö

Contents

Abbreviations

ANN	Artificial Neural Networks
BOW	Bag-Of-Words
CTM	Correlated Topic Model
DTM	Dynamic Topic Model
EM	Expectation maximization
HDP	Hierarchical Dirichlet Process
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
MAD	Median Absolute Deviation
MT	Machine Translation
NLP	Natural Language Processing
PAM	Pachinko Allocation Model
pLSI	Probabilistic Latent Semantic Indexing
RNN	Recurrent Neural Network
SEM	Search Engine Marketing
SEO	Search Engine Optimization
TF-IDF	Term frequency - Inverse document frequency

1 Introduction

Since the beginning of the Internet, one of its primary uses has been enabling communications between people. Usenet groups and newsgroups were used to broadcast news and discussion. The IRC protcol developed in 1988 enabled chat like conversation between individuals and large groups of people. Bulletin boards, chat applications such as MSN messenger and ICQ and later Facebook, WhatsApp and the likes have gained huge audiences during the 2000s.

Public Internet discussion boards provide a view into people's opinions and thoughts that is available for anyone. This thesis aims to find, whether current trends or events can be identified from public Internet discussion data. The motivation for this research is the potential of applying trend information to marketing purposes. Marketing has developed into a more and more data driven direction in the recent years. Extending information inputs from market research and owned data into more complicated external signals has become viable, as technologies and the industry mature.

Traditionally, customer insight for business and marketing use has been produced using questionnaires, customer panels and expert insight. These methods may be costly and do not necessarily extend into particulary broad audiences. Media tracking and manual research of social media feeds, blogs and other external data sources can provide answers to which trends are on the rise or what issues are important for an audience. Finding the signals from large masses of text data without manual labour, can provide an advantage for users of market insight information.

Methods of text mining, and more recently text analytics, have been researched in the context of machine translation, information retrieval, semantic and sentiment analysis and topic modelling for a long period of time, starting from the early 1940s. The rise of accessible and abundant computational resources and software combined with the methods of text analytics have made it easier and easier to develop applications. Topic modelling is the main text analytics method reviewed and applied in this thesis. It is used for unsupervised classification of text data into its underlying or latent topics. The decomposition of text into topics turns unstructured text data into a usable format, ready for further analysis. The most commonly used topic modelling method is the *Latent Dirichlet Allocation* (LDA), which was developed by Blei et al. (2003) and it is also used in this thesis.

The research questions are:

1. **Can current trends or events be identified from public discussion text data?**

2. **What applications the trends have in marketing?**

To answer the research questions, topic modelling methods are reviewed and a method for unsupervised trend identification from discussion data is presented. The methods are applied on discussion data from two Finnish Internet discussion forums: `www.vauva.fi` and `www.suomi24.fi`. Both forums have a large user base, and `suomi24.fi` is the most visited Internet discussion forum in Finland. Since trend

identification, in the context of this thesis, is about finding signals of short term topical concentrations, the trend identification method presented in this thesis draws from the methods of outlier detection. The developed application builds on pre-existing open source software and method implementations presented in peer reviewed papers and published on open source platforms, such as the *Comprehensive R Archive Network* (CRAN) and public *Git* repositories. The application can be reproduced without commercial software. The results of the topical trend identification method are used to answer the first research question.

The results are readily applicable in e.g. product and marketing planning. Automated production of trend information also provides novel uses in marketing automation: automatic generation of search engine marketing keywords, search engine optimization of websites based on trending issues, generation of advert creatives (artistic material used in advertising, e.g. photographs, drawings, or video) and content recommendation, to name a few. These possible applications of the topical trends are studied and used to answer the second research question. Future prospects of developing the trend identification method, as well as the applications for marketing, are studied. Enhancements of the method and application on more domain specific text data sets are likely to provide even more relevant results.

The author and the thesis advisor were employed at the Finnish media agency Dagmar Oy during the process of writing this thesis. The data collection was done in cooperation with Dagmar employees.

This thesis is structured as follows. The history of text analytics and methods of topic modelling, evaluation and visualization are presented in Section 2. Short review of applications of text analytics in marketing, from the viewpoint of both scientific research as well as current commercial applications, is also presented in Section 2. Section 3 presents the application of topical trend identification. Data acquisition and preprocessing are described. The topic models and the trend identification method are presented in Subsection 3.5. Results of applying topic models on Internet discussion data, identified trends and a study of marketing applications of the identified trends are presented in Section 4. Section 5 summarizes the thesis and provides thoughts on future prospects of topical trend identification. Figure and Table references are numbered relative to the Section and Subsection numbers, e.g. Figure 2.2.2 is the second figure in Subsection 2.2.

2 Topic modelling and text analytics in marketing

Text analytics is a broad term covering methods for drawing information from text and converting text into useful data for analysis. This includes e.g. document retrieval, part-of-speech tagging, topic modelling, sentiment analysis and others. Topic modelling is a text analytics technique for extracting *latent* or hidden themes from text. This allows examining a set of documents at a broader level, document indexing, document grouping and semantic analysis.

The use of the trend identification application results for marketing are considered in this thesis, which is why a short summary of the subject is called for. Marketing is managing an exchange relationship between parties: creating and satisfying customers. A very basic and traditional framework of marketing mix divides marketing into product, place, pricing and promotion, the four P:s (McCarthy, 1964). In contemporary sense the marketing function of a firm is responsible for the promotion part of the marketing mix: communicating and advertising the product, the price and the place to the customers in a favourable way through the use of owned or bought media channels such as a newsletter, TV, printed newspapers or digital media e.g. web site advertisement banners or search engine ads. Text analytics is used in contemporary marketing as a source for data for automatically deciding marketing actions, as a tool for assessing customer experience, for understanding customer needs and so on.

2.1 History of text analytics

Text analytics or sometimes *text mining* has its roots in the diverse and interdisciplinary field of *Natural Language Processing* (NLP), which is considered to have emerged in the late 1940s. As the computational resources available for NLP increased and the volume of data available for analysis grew larger, the terms text analytics and text mining were coined to cover the conversion of textual data into useful information in the late 1990s.

The field has developed within many disciplines, which is illustrated by the multiple names for NLP. While NLP is the term used by computer scientists, the field is often called *Computational Linguistics* in linguistics, *Computational Psycholinguistics* in psychology and even *Speech Recognition* within the field of electrical engineering and signal processing. The following paragraphs in this subsection are mainly based on Jones (1994) and Miner et al. (2012).

The first basic ideas related to NLP were information retrieval, speech-to-text transformation and automatic translation or Machine Translation (MT) as it was then called. Patents on early methods for automatic translation were filed as early as the 1930s (Hutchins, 1997). However, the ideas did not reach popularity then. Some NLP related ideas were brought to public knowledge not in scientific writing but in a magazine, *The Atlantic*. In his column Bush (1945) theorised imaginary machines capable of converting spoken language into text to replace manual typewriters as well as a personal data storage called *Memex*, which could automatically associate

user inputs to articles, correspondence, newspapers etc. stored on microfilm in the device. Natural language processing was intertwined with the early development of artificial intelligence. In 1950, Alan Turing (1950) proposed the Turing test to measure machine intelligence through its ability to communicate with a human.

The idea of machine translation was made famous in a memorandum by Weaver (1955) originally written in 1949, where Weaver recorded his idea of languages being similar to ciphers. If the code, which is the translation rules of a pair of languages, could be broken, the text could then be automatically translated. Similar ideas were later made famous by the probably most cited linguistic Noam Chomsky (1957, 1956), who developed the basis for transformative generative grammar. Later, the NLP for syntactic analysis was largely influenced by Chomsky's work. The first international conference related to NLP was the Conference on Mechanical Translation in Massachusetts Institute of Technology held in June 1952 and a related journal, Mechanical Translation began publishing in March 1954 (Reifler, 1954). During the same year, automatic translation from Russian to English was pioneered in the IBM-Georgetown experiment where sixty Russian sentences were translated to English using grammar rules and a simple lexicon (Dostert, 1955). Even though machine translation was an important part of NLP research in its earliest phases, also document indexing was extensively researched in 1950s and 1960s. A great example of developments in those times is the KWIC algorithm by Luhn (1958). Human computer interaction was pioneered by the *ELIZA* NLP computer, which was one of the first programs to attempt the Turing test (Weizenbaum, 1966). During the fifties and early sixties, advancements addressing the problem of machine translation, automatic information retrieval and rudimentary text summarization were made despite the extremely limited computational resources and tools available at the time. Work related to text syntax, semantics and morphology (the forming and structure of words) was also done. The period of flourishing research and results ended as the U.S. government appointed Automatic Language Processing Advisory Committee or ALPAC recommended in its report in 1966 that more research on alternatives for machine translation was needed and effectively stated that the past developments in machine translation were not promising. This lead to funding cuts and halted MT research.

In the 1970s the use of a "bag-of-words" representation was popularized and used in novel information retrieval methods such as the vector space model for automatic indexing by Salton et al. (1975). While early phases of NLP were much about hand-built rule-based systems based on strict grammar, much influenced by Chomsky and grammar theory developed by linguists, the availability of computing resources and development of practical machine learning provided new directions for natural language processing in the 1980s. This period has been called a "statistical revolution", meaning that the hard coded set of rules and assumptions of strict grammar were accompanied by an alternative of a set of probabilistic assumptions less based on linguistics (Johnson, 2009). Text mining methods such as document categorization emerged (Hayes and Weinstein, 1990). While the machine learning approach was first mainly supervised, requiring annotated and tagged corpora, semi- and unsupervised methods also began to flourish. An example of unsupervised

natural language processing is topic modelling which is covered in Subsection 2.2. Recently, ever increasing computational resources and massive data provided new directions for NLP through *artificial neural networks* (ANN) in applications of e.g. translation, generation and classification (Mikolov et al., 2010; Graves et al., 2013).

2.2 Topic modelling

A collection of text is called a *corpus*. A corpus comprises multiple pieces of text called *documents* which can be e.g. articles in a journal, books in a library or posts on social media. Each document is made up of individual *terms* which are usually words, or multiple words seen as a single unit. In text analytics, the text corpus is often viewed as a "bag-of-words" -model (BOW), where the order of the words in a document is discarded and the documents are represented only by the frequencies of the terms in the document. The dimensionality of the corpus is essentially the number of the terms. The intuition behind topic modelling is that a piece of text exhibits multiple different semantic or meaningful topics or themes to the reader. *Topic modeling* reduces the dimensionality of a text corpus into a set of meaningful topics (Blei et al., 2003). While document clustering can be thought of as a wider scope of methods for unsupervised document analysis, topic modelling is focused on finding structure in text through representing a text corpus by a lower dimensional set of topics. Topic models can be used e.g. to study the underlying themes of a corpus, to classify and annotate the corpus documents based on the topics or to organize, summarize and search the documents. Another approach to topic modelling is *topic segmentation*, which tries to distinguish segments of topically coherent text within a *single* document (Purver, 2011).

Automatic information retrieval The basis of topic models lies in the field of information retrieval, which is finding objects from data based on a query such as a list of keywords (Deerwester et al., 1990). A basic and widely used model for automatic document indexing and information retrieval is the classic vector space model by Salton et al. (1975) which relies on the so called term-frequency inverse-document-frequency or TF-IDF strucutre of a corpus. There, each term in a document is given a value based on the term frequency in the document multiplied by the inverse of the document frequency or in how many documents the term occurs (Sparck Jones, 1972). Comparing the TF-IDF indexes of a document and a query via e.g. cosine similarity, an output of how far the document is from the query is obtained. Using this, the closest documents to a query can be returned. The classic vector space model has a serious drawback: the problem of *synonymy and polysemy*. Synonymy means that two different terms might have semantically same meaning such as *forest* and *woods*. Polysemy means that a single term has multiple semantic meanings based on context such as *book* meaning both the object that can be read and the verb "to book" meaning to reserve a hotel room. In automatic information retrieval a model that cannot deal with synonymy and polysemy fails e.g. when the user queries for "car" but the documents contain only "automobile" (synonymy) or when the user queries for "surfing" the retrieved documents contain "Internet" instead

of the sport "surfing" (polysemy) (Deerwester et al., 1990).

Latent Semantic Analysis, LSA Developed in late 1980s, latent semantic analysis (LSA), or latent semantic indexing (LSI) when used in the setting of information retrieval, was able to overcome the problems of synonymy and polysemy in automatic information retrieval (Dumais et al., 1988; Deerwester et al., 1990). Instead of relying on the straightforward term structure and the TF-IDF index of the document, LSI is based on finding *latent* or *hidden* semantic structure of the document. The documents and queries would then be represented not by the terms themselves, but by the *concepts* the terms refer to. Before LSI and even before the TF-IDF vector space model, the latent structure of documents has been researched through hierarchical document clustering (see (Baker, 1962; Jardine and van Rijsbergen, 1971)) and factor analysis of documents (see (Borko, 1963; Atherton and Borko, 1965; Ossorio, 1966). LSI approaches the problem again from the direction of factor analysis but with far greater computational resources than before. A singular value composition (SVD) is performed on the document-term matrix and the components with largest variance are retained. Omitting the small components reduces the original document-term matrix into a lower dimensional representation of the corpus. The model essentially produces a set of orthogonal factors and each term and document can be represented as a vector of its factor values. The reduced model still retains the possibility to compare term and document vectors in the lower dimensional space to each other by their cosine. The factors can be interpreted as the latent *topics* of the corpus. Since a single term or a single document can be mapped to similar factor value vectors, the problems of polysemy and synonymy are less of a problem compared to the classic vector space model.

Probabilistic Latent Semantic Indexing, pLSI LSI deals adequately with synonymy and polysemy but the explicit synonym usage of words cannot be uncovered. In addition, LSI does not provide a sound probabilistic model for evaluating the *fitness* or goodness of fit of the latent topics. It is useful to study, whether LSI is able to capture elements of a generative model on a corpus (Papadimitriou et al., 1998). Proabilistic latent semantic indexing (pLSI or pLSA) is an alternative approach to the same problem as LSI, but with an underlying generative probabilistic model which is fitted to the observed data using expectation maximization (EM) algorithm (Hofmann, 1999). pLSA models the corpus as a *mixture model* where each term in a document is a sample from the model, where the mixture components are multinomial random variables which can be interpreted as the latent topics. The only hyperparameter in the model is the number of topics k which is set before the model is fitted. The terms are not observed through the documents themselves but through the latent topics:

- Select a document d with probability $P(d)$.
- Pick a *latent class* z with probability $P(z|d)$, which is multinomially distributed.
- Generate a term w with probability $P(w|z)$, which is multinomially distributed.

The observed result is the pair (d, w) and the latent class z is not observed. Each term is generated from a single topic and different terms in the document may be generated by multiple topics. Each document is represented by a fixed mixture of topics. The generative model is visualized in Figure 2.2.1. The model has two assumptions: Firstly, the generated pairs (d, w) are independent. The second assumption is conditional independence of the words w from the documents d; the words are only conditioned on the latent class z. The model is fitted by maximizing the likelihood of the data given the model using the EM algorithm: first the posterior probabilities of the latent classes are computed from the current parameter estimates (the expectation or "E" step) and then the parameters are updated to maximize the posterior probability (the maximization or "M" step).

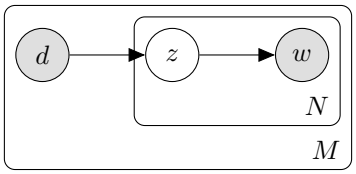


Figure 2.2.1: Plate diagram of the generating process of the pLSI model. There are M observed documents d, which are represented by the mixture (a multinomial distribution) of latent topics z, from which N terms w are drawn for each document. The illustration is similar to what is presented by Hofmann (1999).

pLSI was a great leap over the previous factor analysis model but it has some shortcomings: Since the documents are represented by fixed mixtures of topics without a generative model, the number of parameters grows linearly with the number of documents which can lead to overfitting. Secondly, the model does not provide ways to assign topic probabilities to documents outside the training corpus.

Latent Dirichlet Allocation *Latent Dirichlet Allocation* (LDA) is a two level, generative probabilistic model of the corpus. LDA essentially extends pLSI with a generative model for the documents. Developed by Blei et al. (2003), LDA is the industry standard and very widely used technique for topic modelling. LDA retains the structure of pLSI where each document is represented by a mixture of topics and the topics are represented by a mixture of terms. The meaning of the latent topics is interpreted from the term-topic -mixture and is covered in Subsection 2.3. The development of LDA paved the way for a set of more complex topic models that share the same basic document-topic-mixture and term-topic-mixture structure, but extended the capabilities of the model in other ways.

LDA replaces the fixed topic distribution of pLSI by a generative process, where the topic mixture θ of a document is drawn from a Dirichlet distribution parametrized by α. The Dirichlet is a distribution over a simplex, meaning that the drawn vector sums to one. Apart from modelling the topic distribution of the documents with a Dirichlet instead of a static topic mixture, LDA is similar in structure compared to pLSI as follows:

- Sample the number of terms N in a document from a Poisson distribution.

- Sample the multinomial topic distribution θ in a document from a Dirichlet distribution parametrized by α.
- For each N term w_n:
  - Sample the topic z_n of the term from the multinomial distribution of topics θ.
  - Sample the term w_n with probability $p(w_n|z_n, \beta)$, which is a multinomial distribution conditioned on the topic z_n.

The LDA model is visualized as a plate diagram in Figure 2.2.2. As with pLSI, the number of topics k is the only hyperparameter of the model and it is fixed before fitting the model. Since LDA has a complete generative model, choosing k can be done by calculating the model likelihood on held out data or by evaluating the interpretability of a model given k. This is covered in Subsection 2.3.

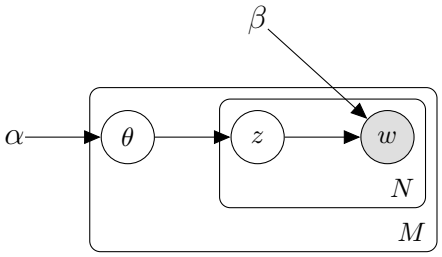


Figure 2.2.2: Plate diagram for the generating process for Latent Dirichlet Allocation. There are M documents and the topic distribution θ of a document is sampled from a Dirichlet parametrized by α. In each document, there are N terms which are generated by sampling a topic z from θ and then sampling the term from the topic's multinomial term distribution β conditioned on z. The illustration is similar to what is presented by Blei et al. (2003).

Fitting an LDA model is essentially inferring the posterior probability of the model, which is the joint probability over the terms, documents and topics. There are multiple algorithms for inferring the posterior such as mean field variational Bayesian methods (Blei et al., 2003), collapsed Gibbs sampling (Griffiths and Steyvers, 2002), collapsed variational inference (Teh et al., 2007) and online variational inference (Hoffman et al., 2010). As data set sizes have increased in recent times, the need for faster and more memory efficient inference methods for LDA has risen. New methods deal with the problem of term sparsity as the number of terms in the corpus increases as well as splitting the workload for multiple, even thousands of compute instances (Newman et al., 2009; Yao et al., 2009; Yu et al., 2015). LightLDA is an inference framework for LDA developed with smaller computing resources in mind (Yuan et al., 2015). An even more promising memory efficient inference method called *WarpLDA* has been developed by Chen et al. (2016). WarpLDA is the inference method used for LDA in the application of this thesis.

LDA has been applied in multiple fields and uses, even outside text analytics. This list is by no means complete but it illustrates the wide range of possible uses:

- Text analysis and document collection analysis (Boyd-Graber et al., 2007; Hall et al., 2008; Jockers and Mimno, 2013)
- Analyzing microblog services (Ramage et al., 2010)
- Satellite imagery annotation (Lienou et al., 2010)
- Internet content tagging recommendation (Krestel et al., 2009)
- Social network recommendation (Chen et al., 2009)
- Identification of gene function and genetic functional modules (Liu et al., 2010; Pinoli et al., 2014)

The last examples of the applications of LDA in genealogy are interesting, because a very similar model to LDA was previously and independently developed for genetics (Pritchard et al., 2000).

Variants and applications of LDA Extensions of the basic LDA have been developed to augment the capabilities of the model for various tasks. Possibly the second most popular topic model currently used, after LDA, is the Correlated Topic Model (CTM) by Blei and Lafferty (2005), the same group that originally published LDA. The underlying Dirichlet distribution of topic mixtures in LDA assumes that the components are independent. Thus it follows that LDA is unable to capture the possible links between different topics. If a discussion forum post exhibits a topic "relationships" it is more likely to exhibit also a topic "family" than "steam locomotives". CTM replaces the Dirichlet with the logistic normal distribution, which, again, is a distribution on the simplex but it is able to model the dependence between the components. In the generative process of a CTM, the log of the multinomial distribution parameters are drawn from a multivariate normal distribution. The model is visualized in 2.2.3.

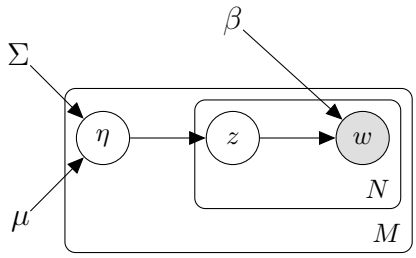


Figure 2.2.3: Plate diagram for the generating process for Correlated Topic Model. The Dirichlet has been replaced by multivariate logistic normal distribution to allow dependencies between the topics. The illustration is similar to what is presented by Blei and Lafferty (2005).

Other approaches to the modelling of the correlation structure between topics exist as well. Pachinko allocation model (PAM) replaces the two way correlations of topics in CTM with an directed acyclic graph (Li and McCallum, 2006). Instead of

a single level of topics and their correlations, PAM can model *super-topics* and their *sub-topics* (and even more levels of hierarchy) and a correlation structure between the sub- and super-topics. This allows for a more fine grained structure of the topic model to exhibit both broader subjects and detailed, small topics. Visualization of PAM through a graph and comparison to LDA is presented in Figure 2.2.4. In addition to PAM, the Hierarchical Dirichlet Process (HDP) is an another approach for modelling a grouped clustering problem of data such as the correlation structure of a topic model (Teh et al., 2005). HDP can be applied to topic modeling where it either is used to model a nested correlation structure of topics or to select a suitable number of topics.

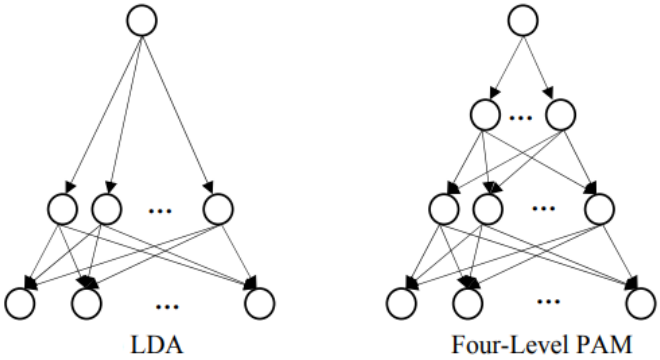


Figure 2.2.4: Comparison of Latent Dirichlet Allocation model (LDA) and a four-level Pachinko Allocation model (PAM), graphics presented originally by Li and McCallum (2006). The top node represents the entire corpus. The leaf nodes in the graph represent individual terms. The inside nodes represent topics. An edge from a node to another represents a connection: A super-topic comprises multiple sub-topics and a sub-topic comprises of a multinomial of all terms. A super-topic may have connection to multiple sub-topics which represents the correlation of topics in PAM.

LDA and CTM model the corpus with a static term-topic mixture. Dynamic Topic Model (DTM) extends LDA to allow the term-topic mixture to vary over time, when the documents in the corpus are assigned into time slices. (Blei and Lafferty, 2006). DTM is a family of different models but in a basic form the natural parameters β and α are generated from a simple Gaussian state space model. Inference of such model can be done by using e.g. Kalman filter or wavelet regression based variational methods. DTM can be successfully applied to data where the contents of a topic evolve over time. In the original paper, the topics of the journal *Science* are successfully modelled in a time period from 1880 to 2000 such that each year is a single time slice with its own term-topic mixtures. In the results, a topic that can be interpreted to be about "Atomic physics" exhibits the term "matter" much more in the early years compared to present time and the term "quantum" in an opposite trajectory with little significance in 1880 and large importance in present time. A dynamic topic model for continuous timesteps has also been proposed (Wang et al., 2008).

Dynamic topic models allow the time stamp of a document to affect the topics. An even wider view to the effect of external factors (such as document publication year in DTMs) is provided and modelled by the Structural Topic Model (STM) (Roberts et al., 2014). In an STM, some metadata of the document such as the author, the political view of the author or the journal, is allowed to affect either the term-topic mixture or the document-topic mixture. This enables studying the effect of external factors on the topics.

The previous extensions of LDA provide ways to capture complex structure of a corpus. In some cases, however, the corpus is in fact the opposite and consists of short documents with little information or rich structure. This is often the case with social media related text data, which is an important data source in the field of marketing. Bigram Topic Model (BTM) is a type of probabilistic topic models suitable for examining corpora with few terms per document (Yan et al., 2013). A biterm is an observation of two terms occurring in the same document. The BTM generative process assumes that each biterm is drawn from a term-topic mixture such that the two terms of the biterm are drawn from the same topic. Instead of generating a topic mixture individually for each document, BTM has a common topic distribution for the entire corpus. BTM tries to uncover the latent topics not by looking at terms that occur together in a specific document but rather occurring together in general.

Neural networks and topic modelling There is also research on topic models based on neural networks, such as replicated softmax layers and Boltzmann machines (Srivastava and Salakhutdinov, 2012; Hinton and Salakhutdinov, 2009). Recurrent neural networks are able to capture dynamic relationships in a data. *TopicRNN* is proposed application of RNNs to topic modeling (Dieng et al., 2016).

A novel approach to NLP is the use of *embeddings* learned by (the use of) neural networks. A word embedding is a high dimensional vector representation of words, learned from a data. It allows the study of the relatedness of words to another as well as "summation" of words (e.g. "king"-"man"+"woman"="queen"). A famous neural network based word embedding method is the *word2vec* (Mikolov et al., 2013). Building on the word embeddings of *word2vec*, *lda2vec* is a recent topic modeling and word embedding method for learning both word embeddings as well as the LDA topics. The model predicts words in a document, not only by topic distributions but also the word embedding (Moody, 2016).

Supervised topic models While topic models are specifically created for finding the latent classes in a data set (or topics in a corpus), sometimes *supervised* analysis of a text corpora is useful. Supervised topic models are useful for prediction tasks: movie rating prediction based on reviews, topic labelling or document tagging. Some models assume a single response variable on each document in a training set (Mcauliffe and Blei, 2008; Lacoste-Julien et al., 2009) while other enable a more complex label structure of multiple labels with different weights (Ramage et al., 2009).

The list of topic models presented in this subsection is far from complete and multiple further variations can be found in the literature. An overview to probabilistic

topic models such as the ones covered in this subsection is provided in Blei (2012).

2.3 Interpretation and visualization of topic models

Probabilistic topic models provide a latent topic representation of a text corpus. Each topic is essentially a distribution of word occurrence in a topic and each document in the analysed corpus is modelled as a distribution of topics. Since the number of topics can be large and the used vocabulary even larger, up to the order of even 100 000 terms, evaluating the fitness of the model, interpreting the meaning of each topic and visualizing the results is a substantial task.

2.3.1 Evaluating model fit and interpreting topics

The semantic meaning of the latent topics is not known or fixed beforehand. Thus, after the analysis, careful examination must be done to find out whether the found topics really have a semantic meaning and what it is. Important aspects related to topic interpretation are defining quantities on term importance in a topic and explicit topic labelling. The quality of the model fit can be assessed from the viewpoint of calculating probabilistic model fitness for held-out data, evaluating the coherence and intepretability of the model results using human or automatic evaluation as well as by identifying "bad topics" in the model.

Term importance quantities The word-topic distribution usually has high probabilities for words that represent the main concepts of each topic. Examination of these high probability terms is often used to empirically determine the meaning of each topic (Blei et al., 2003; Blei and Lafferty, 2005; Chaney and Blei, 2012). While probability is often used as the metric for determining the most important terms of a topic, other metrics exist as well. A measure called *lift* can be used to highlight terms that are specific to a single topic. It is defined as the ratio of term probability within a single topic and the term's marginal probability across the entire corpus (Taddy, 2012). However, problems with noise may arise and terms that are very rare and occur only within a single topic may receive disproportionately high ranking. *FREX* or FREquency and EXclusivity is a harmonic mean of term frequency and exclusivity (the proportional frequency within a topic compared to a set of comparison topics) (Bischof and Airoldi, 2012). Term *distinctiveness* quantifies the amount of information a term conveys about a topic. It is defined through the Kullback-Leibler divergence between the distribution of the topics given a term and the marginal distribution of the topics given a term. Distinctiveness weighted by the term's frequency in the entire corpus is called *saliency* (Chuang et al., 2012a). Saliency can be used to effectively select the terms that are relevant in interpreting topical meaning. Building on top of these metrics, Sievert and Shirley (2014) combine term frequency and lift in a metric called *relevance*, which is defined as the weighted average between log frequency and log lift. The weighting factor was tuned based on the results a user study. The relevance metric is used in the application of this thesis.

Topic labelling Even though probabilistic topic models are designed for finding latent topics in the corpus, it would be very useful to uncover the explicit meaning of each topic automatically. A method proposed by Mei et al. (2007) searches for frequently occurring chunks of text in a topic and the one that has minimal Kullback-Leibler divergence with the word distribution of the topic and maximal mutual information with the topic is selected as the label for the topic. The method has been used for different types of text chunks e.g. n-grams and part of sentence (POS) based chunks. Maximal uniqueness of a topic label with respect to other topics can also be taken into account. Another approaches to automatic topic labelling are based on learning topic labels from Wikipedia titles (Lau et al., 2011) and Twitter (Zhao et al., 2011). Related to supervised topic models, there are also methods for setting a priori constraints on specific topics. This may be helpful in interpreting the explicit meaning of the resulting topics (Andrzejewski et al., 2009).

Model evaluation methods If semantic meaning is cast aside, evaluating the fitness of a topic model is traditionally done through calculating the log likelihood for held-out documents (Wallach et al., 2009). In practise, it means that after training the topic model on a portion of the entire corpus e.g. 80%, the word probability of the remaining (held-out) documents given the trained model is estimated. Wallach et al. (2009) introduce multiple methods for held-out document likelihood estimation. Instead of the model log likelihood, another metric called *perplexity* can be used. It is the exponential of the negative log likelihood scaled by the number of tokens in the corpus. Perplexity is suitable for comparing the fitness of models with different data sizes.

While the held-out document likelihood is a useful metric for model and model parameter (such as topic number) selection, it does not provide information about the semantic meaningfulness and interpretability of a model. One approach to evaluate the meaningfulness and usefulness of the latent topic space is human evaluation. Chang et al. (2009) present two metrics and a formal test setting for topic model evaluation: word intrusion and topic intrusion. The two metrics help to evaluate whether the topic model is able to match the human concepts or does it give unmatched terms a high probability in a topic (word intrusion) and whether the topic model and a human agree on the topic assignments on a document (topic intrusion). The results show that these validation metrics based on human experiments correlate negatively with the traditional held-out document likelihood metric of the model (Chang et al., 2009). This suggests that more emphasis on real-world task performance rather than held-out document likelihood should be put in topic model development. On the other hand, the same results provide sound base for the assumption that topic models and humans agree on the semantic coherence of topics and documents' topic assignment.

Other topic interpretability metrics have been formalized based on *Wikipedia*, *Google* and *WordNet* -resources. It was found that e.g. a metric based on matching words from 100 first Google search results on the document title to the first 10 words in the topic word distribution performed nearly as well as the so called inter-annotator agreement metric produced by human annotators (Newman et al., 2010).

This suggests that automatic topic model interpretability analysis could be achieved using non-human metrics. Further research in automatic topic model evaluation has been conducted (Chuang et al., 2013; Alexander and Gleicher, 2016).

In some cases, a portion of the topics generated by the model are interpretable and coherent to human reader, but some are not. One approach to identify non-meaningdul topics is based on the loosely defined *Zipf's law of words*, which states that in a genuine topic, the word distribution is skewed towards a small set of important or *salient* words of the entire corpus rather than a uniform distribution of all words. Using this as a basis, multiple decision criteria regarding the word frequency distribution in topics can be formalized and used to distinguish bad topics from the meaningful ones. (AlSumait et al., 2009)

Another method for identifying "bad topics" is based on four categories of "bad topics" which have been identified from human expert evaluations. A coherence metric that draws from the insight on these categories can then be used to identify "bad topics" and overall model interpretability performance. Comparing the results from the coherence metric and expert opinions on topic quality shows a good connection. In tandem with the coherence metric a *Generalized Pólya urn* topic model was developed to prevent bad topics from forming in the model. (Mimno et al., 2011)

2.3.2 Visualizing topic models

When a model has been trained and validated to produce meaningful topics, it becomes relevant to visualize topic model information to a user in a possibly interactive way. The structure of probabilistic topic models, the word-topic and the document-topic mixtures, forms the basis for visualizing topic models.

A visualization method developed by the original authors of the popular LDA model and drawing from previous research (Gardner et al., 2010) is shown in Figure 2.3.1 and accessible in `http://www.princeton.edu/~achaney/tmve/wiki100k/browse/topic-presence.html`(Chaney and Blei, 2012). It is a basic but effective way to browse the topics and their word distributions, the related topics of a topic (based on word distribution similarity), the documents related to a single topic, the topics related to a single document and the topics related to a single word. The three sides of the model - the corpus, the words and the topics - are all interactively accessible in a meaningful way. Other visualization ideas are also present in this framework: Word probability in a topic is displayed via the use of a horizontal barplot and a pie chart is used to present the topic probabilities of a document. Each topic is represented by a list of three most probable words in the topic.

When considering the visualization of only a single topic instead of e.g. the interaction of multiple topics, a word cloud is a basic and functional method. It is often used when visualizing a text corpus or a topic model (Gardner et al., 2010; Ramage et al., 2010; Jockers and Mimno, 2013). A number of important terms for a topic (or in general, any piece of text) are selected and drawn in a figure such that the font size for each word is proportional to a metric describing the importance of the word, such as frequency. The words are often laid out so that they fit together tightly and the words that have larger font i.e. importance, are placed in the center.

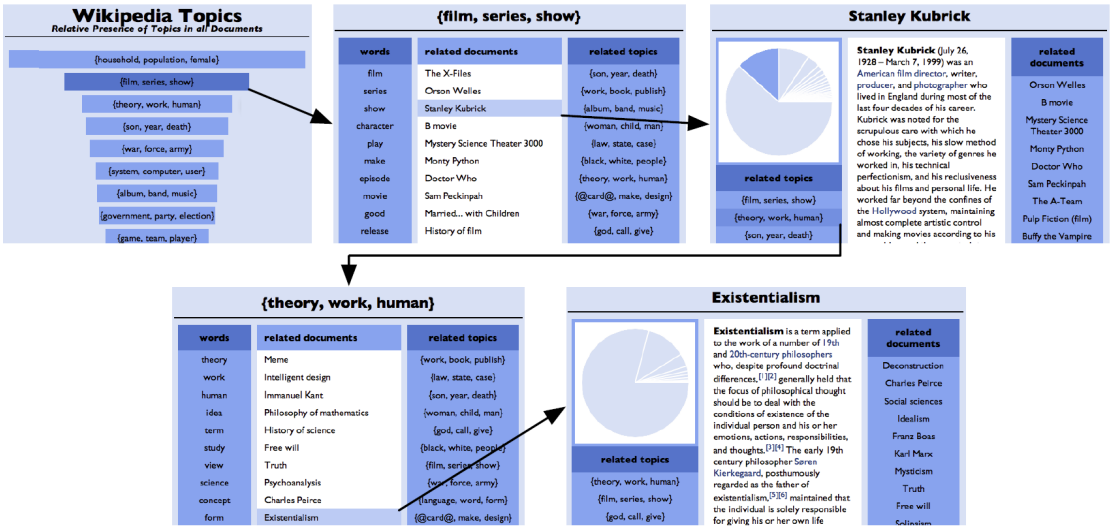


Figure 2.3.1: Example screen captures of a visualization tool for topic models. The visualization tool is accessible in http://www.princeton.edu/~achaney/tmve/wiki100k/browse/topic-presence.html, accessed 2.11.2018. Rotating from top left following the arrows the images represent the following: 1: A view of all topics sorted by topical volume and labelled by the top three most frequent terms of the document. 2: A view of a topic showing its top words, related documents and related topics that have similar term distributions. 3: A view of a document showing its original contents, related topics and topic proportions and related topics that have similar topic distributions. 4: Another topic view. 5: Another document view.

An example of a word cloud based on data used in this thesis is presented in Figure 2.3.2.

The relatedness of topics can be quantified in multiple ways. In the visualization tool shown in Figure 2.3.1, the relationship between topics is calculated pairwise for every topic as the average log odds ratio of the probability of each term in the topics. Another way to calculate univariate topic similarity is the angle of the term-topic vectors (multinomial term-topic mixtures) of two topics (Chuang et al., 2012b). A multivariate similarity can be derived based on these pairwise univariate measures of similarity. Performing principal component analysis or multidimensional scaling on the distance matrix reveals multidimensional similarity structures between the topics. This allows for the user to interpret, which topics are related and which are not (Chuang et al., 2012b; Sievert and Shirley, 2014). However, the selection of the model can affect the visualization result and therefore the interpretation of the model. Thus, it is advisable to use visualization systems that are model agnostic to allow trustworthy interpretation (Chuang et al., 2012b).

Termite is a tool for analysing topic composition and also to visualize topic relationships (Chuang et al., 2012a). The tool shows multiple model metrics: The word frequency on the document-term matrix, which allows easy comparison of similar term mixtures for different topics. The tool also shows the relation of a single topic's word frequencies to the frequencies of the entire corpus. This provides a way

Figure 2.3.2: Example of a word cloud. The words with more importance are drawn with larger font. This example is from `vauva.fi` discussion forum data used in the application of this thesis. The words are stemmed.

to visually interpret, which terms are especially important for the topic. Lastly, the tool shows a set of documents with good topical fit.

LDAvis is a tool for visualizing both two-dimensional topic relations and relevant terms for each topic (Sievert and Shirley, 2014). The tool allows the user to change the weighting parameter in the relevance metric described in 2.3.1 and visualizes both the lift and frequency components in the term list. Lift is visualized in a similar manner by Chuang et al. (2012a). The tool also visualizes the conditional term frequencies in each topic as the user hovers mouse pointer over the terms.

Sometimes the documents in the corpus are timestamped, as is the case in the application part of this thesis. This makes the topic proportions of the corpus to have a dynamic component. Although not a published work, Goldstone (2016) gives an excellent example of a simple tool for visualizing topic models from multiple angles, including time series visualization. Occurrence of documents related to a topic is visualized through the use of a bar plot. The topic proportions over time are visualized in a stacked area chart. The tool also includes visualization of topic similarity in two dimensions, source documents and topics related to a single word.

Visualizing topic models is not only about interpreting the results and meaning of the topics, but also providing ways to visually compare the performance of two different topic models. A *buddy plot* is a novel type of visualization, where the distance of a document to all other documents in the corpus is encoded in location for one topic model and color for another topic model. If the color scale is smooth in the buddy plot, the two topic models provide similar topical representation of the document. The buddy plot and other model comparison visualizations are presented

by Alexander and Gleicher (2016).

Visualization can also be combined with interaction. Tools for interactively visualizing and giving feedback, refining the topics and changing the model have been developed, see e.g. (Lee et al., 2012; Hoque and Carenini, 2015). Recently, also more advanced visualization techniques and frameworks for topic models have been developed, see e.g. (Liu et al., 2012, 2014; Cui et al., 2014).

2.4 Applications of text analytics in marketing and media

Availability of text data through the Internet has enabled the use of text analytics for marketing. Text analytics is used to provide insight on customer behaviour and competitor activities, automation of marketing activities and even content creation. Analytics for marketing in general organizes unstructured data into information that is useful for business decisions. Rackley (2015) suggests a practical approach on applying analytics for marketing. While analytics in marketing have been studied and tools for tracking relevant metrics for marketing are available, analysing textual data is a less researched subject. However, numerous commercial applications and software for text analytics for marketing have emerged. Lists and reviews of applications are readily available on the Internet e.g. `https://www.predictiveanalyticstoday.com/` (accessed: 7.12.2018). Tools for text analytics in marketing provide means for content search and tracking, content categorization, sentiment analysis of content, chatbots, discussion moderation and search keyword generation.

Media tracking is mining media content on the web and extracting information related to relevant keywords, such as brand names, people or events. An example of media tracking software is a Norwegian media tracking platform *Meltwater*, which allows users to track media hits on keywords. Based on text analytics, the platform returns the number and context of media hits. Other large media tracking platforms include *Cision media tracking* and *Mention*.

Sentiment analysis is calculating the *sentiment* of text or other media content in a context. This helps marketing to assess how a brand or brand related affairs are discussed online. *Mediatoolkit* and *InsightAtlas* are examples of global sentiment analysis platforms. The process often includes media tracking, and sentiment analysis is used to further enrich the information. Even more information on context can be provided with content categorization and keyword extraction. For example, *Leiki* is a Finnish platform for automatically categorizing the content of text. This gives marketers the ability to automatically organize content based on the category and meaning, as well as to understand what is talked about. Automatic text categorization can be based on explicit ontologies (such as *Leiki*) or latent topics, as is done in topic modelling. Sentiment analysis and content categorization give sufficient insight on content to e.g. automatically generate search keyword lists, such as the *WordStream* platform or to even automatically create text content for marketing purposes, such as the *Articoolo* platform.

In recent times, chatbots or automatic text messaging interfaces have been developed for marketing use. Although the subject has not been researched much, indication for readiness of chatbot interaction has been found (Eeuwen, 2017). Chat-

bots apply advanced text analytics and NLP methods to communicate with human users. Chatbots may answer to questions, provide information or employ marketing actions such as present advertisements or purchase suggestions. Chatbots may be either conversational using spoken language such as *Siri* by Apple, *Assistant* by Google or *Alexa* by Amazon, or textual, embedded on a website such as *Giosg* chat interface and chatbots. Another application of text analytics closely related to chatbots is automatic discussion moderation, which also employs NLP techniques. Automatic moderation enables marketers to remove abusive content from discussion at large scale, removing heavy workload from human moderators. *Utopia analytics* is an example of a platform providing automatic discussion moderation. It is in pilot use on a Finnish discussion forum `www.suomi24.fi`, which is used a data source for this thesis.

3 Application of topic modelling in Finnish Internet discussion forums for trend identification

The application of this thesis is to use topic modelling to identify current trends in public discussions. This thesis is was produced in co-operation with the Finnish media agency Dagmar Oy, where the author was employed during the process of writing this thesis. Data used in the application was collected by Dagmar employees.

In order to identify trends and draw conclusions on public opinions and themes in public conversation, the data must represent the thoughts and discussions of a wide group of the target audience, i.e. the Finnish people. The discussion forums `www.suomi24.fi` and `www.vauva.fi` were chosen as the data source, because they fulfil the above mentioned requirements sufficiently. The discussion forum data is free and publicly available, it has large volume and it represents a wide group of Finnish people. The popularity and data volumes of `www.suomi24.fi` and `www.vauva.fi` are presented in the following subsections. The amount of documents i.e. text needed for a topic model to have sufficient performance has not been extensively researched. However, the daily post volumes are nearly 10 000 on the discussion forums in question, which is likely to be more than enough for this application.

The data was preprocessed for topic modelling and Latent Dirichlet Allocation models as well as Dynamic Topic models were fitted in the data to discern the underlying topics of the public discussions. Models were evaluated and tuned according to model perplexity as well as human evaluation. A custom trend identification algorithm was developed to identify topical trend events from the proportional topic volume time series. The trend identification was applied only on LDA model outputs.

3.1 Discussion data description

The data source for the application in this thesis is a collection of internet discussion forum posts from two Finnish websites, `www.suomi24.fi` and `www.vauva.fi`. An internet discussion forum is a social media platform comprising a collection of posts by users with possibly anonymous user names. Forum posts are often arranged in one or more levels of *subcategories*. A new post at a subcategory starts a new *thread* with a title determined by the first post. Other users may reply to the thread with posts. A post may contain text, *emojis*, website links and cited text from other posts. The discussion forum data used in this thesis consists of individual posts as data points with multiple components: post content, thread title, thread URL, forum subcategory and time stamp of post submission. The forum platforms do not allow threads updated far in the past to be accessed, which sets limitations on data acquisition. For example, `vauva.fi` displays the first 400 pages of threads on the website.

suomi24.fi `www.suomi24.fi` is a social media website which was founded in 1.10.1998. Along with other content, such as news aggregation, dating service and link indexing, the website has had a discussion forum since the beginning. During the website's history, it has been owned by different media companies and is now

a private company wholly owned by Aller Media Oy, a Finnish publishing company. `www.suomi24.fi` is the sixth most visited Finnish website with a monthly reach of circa 2.3 million users measured on all different devices in May 2018 (Ourila, 2018). It is the most popular Finnish internet discussion forum by website reach.

vauva.fi Vauva is a Finnish magazine about parenting, pregnancy, baby care and families published by Sanoma Media Finland since 1992. The magazine website `www.vauva.fi` is the most popular magazine website by reach in Finland as of May 2018 with circa 1.6 million users on all different devices reached in May 2018 (Ourila, 2018). The website hosts a popular discussion forum which is the second of the two data sources for this thesis. The discussion forum itself goes by the name *"vauva.fi"* in spoken language. Despite lower user reach than `suomi24.fi`, `vauva.fi` discussion forum has higher post activity, as will be explained in Subsection 3.3. This indicates that at `suomi24.fi`, there are more people reading and following the discussion proportional to active participants of discussion, whereas at `vauva.fi` a larger proportion of website visitors not only read but also participate in the discussion.

3.2 Data acquisition

Discussion forum contents for the application were extracted from the source websites using a web crawler. Data acquisition via a crawler and the content available at the website was preferred over purchasing the data from a vendor or directly from the publisher to ensure future data availability and to validate the data collection methods.

A focused web crawler is a program which indexes web page contents within a certain website. Web crawlers were first used to complement traditional search engines but later have gained use in information retrieval (Ester et al., 2004). The process of automatically extracting data from a website using a web crawler is called web scraping.

The web scraping process used for data acquisition for this thesis was developed by Teemu Huovinen, Dagmar Oy using Python library `scrapy` (Scrapinghub, 2008). The scraping scheme was adjusted individually for both source websites, `suomi24.fi` and `vauva.fi`, to suit the website and forum structure. In both websites, the basic idea for scraping was to sequentially visit each discussion thread found on the forum main page. The thread list was ordered by the timestamp of the latest post with a limited number of threads displayed on the page with the following threads found on the next pages. Within the thread each post was sequentially processed starting from the latest one. When a post with a timestamp older than the time of the previous crawler update run was encountered, the crawler returned and proceeded to the next thread. When all threads on the main page were crawled, the program identified the link to the next page of threads and continued until a thread with no new posts after the previous crawler update was found.The scraping scheme is visualized in Figure 3.2.1.

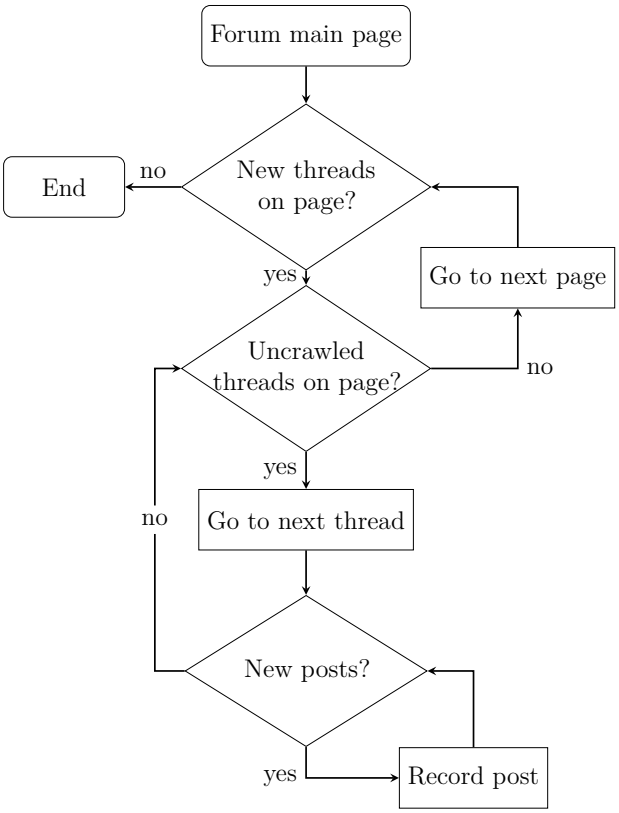


Figure 3.2.1: Web scraping process of internet discussion forums `www.suomi24.fi` and `www.vauva.fi` used for hourly data acquisition update.

After initially crawling the forums for all available threads, the data was updated hourly by running the crawler. Each individual post was recorded in *Google Storage* (Google, 2010) buckets in `.jsonl` JSON Lines file format, which means essentially newline separated JSON objects (Ward, 2017). JSON or Java Script Object Notation is a cross-language text syntax with simple name-value pair data format which is easy to read and parse both for machine and human users (Bray, 2017). The files had data fields listed in Table 3.2.1. JSON Lines files were then uploaded from Google Storage buckets to `R` programming environment and all further analysis was conducted using `R` or `C` programming languages.

Table 3.2.1: Data storage fields JSON Lines files in Google Cloud Storage bucket.

Field name	Description
`category`	Forum subcategory of thread
`comment_text`	Post contents
`thread_title`	Thread title
`timestamp`	Time of post submission
`url`	URL of the thread

3.3 Data exploration

The data crawl began on 7.5.2018. While crawling of discussion extends far into past by accessing threads with long history, complete data was estimated to be accessible only from 1.5.2018 due to the limited number of recently updated discussion threads accessible at the website pages. Discussion forum data used for the thesis application was from the time period of 1.5.2018-30.11.2018. Discussion forum activity was measured with post volume. The total number of individual forum posts was 1 750 720 in `suomi24.fi` and 2 438 593 in `vauva.fi`, during the time period of the thesis application. The hourly, daily, weekly and monthly mean post volumes for both forums are in Table 3.3.1. `vauva.fi` discussion was open from 7.00 to 23.00, meaning that users were not admitted to post outside of the opening hours. The opening hours are taken into account in the hourly mean of post volume in Table 3.3.1. Total daily and weekly post volume time series are in Figure 3.3.1. Scraping of both `suomi24.fi` and `vauva.fi` had problems on a couple of occasions. Outlier data points of `suomi24.fi` scarping are highlighted in Figure 3.3.1. The missing posts were not recovered for final data used in the application. The post volume varies depending on weekdays as well as the times of day. The mean post volumes over weekdays and hours of day during the data period are presented in Figure 3.3.2.

Table 3.3.1: The daily, weekly and monthly mean post volumes for both source discussion forums from time period 1.5.2018-30.11.2018. Hourly mean is calculated only over hours when posting on the forum was admitted.

Post volume	`suomi24.fi`	`vauva.fi`
Weekly mean	53052	73897
Daily mean	7781	10838
Hourly mean	336	637

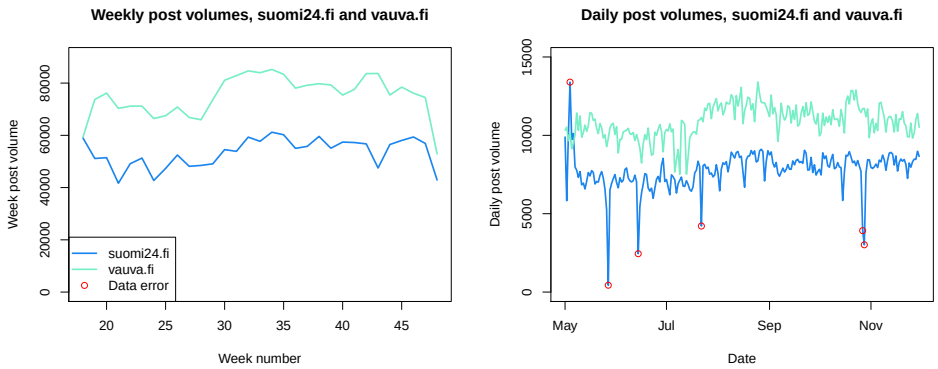


Figure 3.3.1: Post volumes by week and day for `suomi24.fi` and `vauva.fi` from 1.5.2018 to 30.11.2018. Data collection errors resulted in drops in captured post volumes for `suomi24.fi`

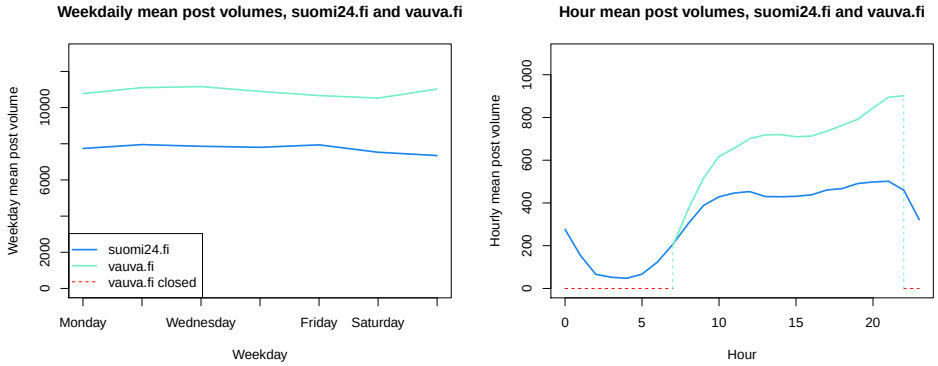


Figure 3.3.2: Post volumes by weekday and hour of the day for `suomi24.fi` and `vauva.fi` from 1.5.2018 to 30.11.2018. Opening hours of `vauva.fi` are marked in hourly volumes.

Problems in website crawling Although a lot of discussion forum posts were recorded during data crawling, it became evident that not every single post was recorded. First indications of problems in crawling were visual outliers in the total crawled post volume plots, e.g. in Figure 3.3.1. Further investigation revealed that HTTP-504 or gateway time out errors were logged in some cases, but no errors in some other cases where it was evident that posts were missing. Re-crawling the time periods of missing posts for `vauva.fi` was able to restore some of the missing posts but not all of them, which can be seen in Figure 3.3.3.

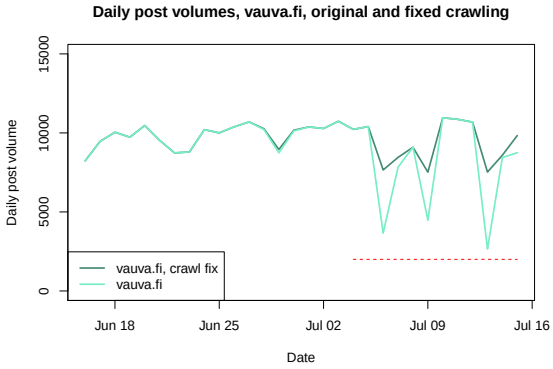


Figure 3.3.3: Comparison of `vauva.fi` daily post volumes before and after re-crawling the time period of missing posts which is indicated by the red dashed line. Not all missing posts were restored.

`suomi24.fi` had also two clear problematic days with missing posts, which are seen in Figure 3.3.1. These time periods were crawled again and more posts were recorded. These problems raised the question of how much posts were missed continuously on days when there was no clear indication of problems in crawling. Since there was no definitely true information about the total post volume available, rate of missed posts was estimated by rerunning the crawler on a six day period of

forum posts for three times. The union of the three crawler runs was assumed to be all posts from that period of time. Then each individual crawler run was compared to the union set and the mean proportion of missing posts was calculated to be only 0.6 ‰. Little could be done about restoring all missed posts, so the number of the posts in the data acquired with single crawler pass was deemed to be enough for the purposes of the application.

3.4 Data preprocessing

Before applying any topic modelling methods or analysis, the unstructured text data extracted from the discussion forums had to be preprocessed into a suitable format. The data preprocessing pipeline followed a scheme suggested by Aggarwal and Zhai (2012) and is summarized in Table 3.4.1.

Table 3.4.1: Data preprocessing pipeline used in the application.

Preprocessing phase	Main goals	`R` package used for processing
Thread title to first post	Adding the thread title as the first part of the first post of each thread, since the title is written by the author.	`data.table`, `R`
Tokenization	Split documents (i.e. forum posts) into lists of individual terms. Tokenization rules include splitting at whitespace and punctuation.	`tokenizers`
Text stemming / lemmatization	Truncate words into their basic lemma using SnowballC -stemming. Reduces the number of unique terms with little difference in meaning.	`SnowballC`
Text placeholders	Decrease the number of unique number and emoticon terms in the corpus.	Vanilla `R`
Stop word removal	Remove little words with no semantic importance. Reduces the size of corpus.	`text2vec`
BOW vocabulary, pruning	Remove terms with too few or too many occurrences from the vocabulary.	`text2vec`
Document term matrix	Calculate term frequencies in each document and generate the document term matrix.	`text2vec`

Tokenization The text preprocessing begins with tokenization or splitting entire text documents (discussion forum post entries in this application) into terms to be used in a bag of words model. Text tokenization is based on rules on how terms are split in the sentence. Typical rules include splitting words at white space or punctuation. Tokenization was done in such a way that only punctuation not related to typical emoticons were used for splitting and emoticons were preserved. Tokenzation was done by using the `R` package `tokenizers`, which allowed fine grained selection of word splits.

Stemming and lemmatization Complex morphology of the Finnish language complicates text analysis based on a bag of words model. Terms with very similar semantic meaning are written differently based on their morphology or case in a sentence such as "ruoka" ("food") and "ruokaa" ("food" as an object in a sentence). Lemmatization is the process where each term is reduced to its *lemma* or the basic "dictionary citation" of the word. In Finnish, for example, the genetive form of the word "food", "ruuan", would be lemmatized to "ruoka". Stemming is a subcategory of lemmatization, a process where each term is shortened from the end until a common root *stem* of the term is reached decreasing the number of distinctive terms in the corpus significantly. Stemming is far more simple process than complete lemmatization, since the entire morphology of a language must not be known. Stemming tool "SnowballC" was used in this thesis to lemmatize the corpus. It was implemented in the `R` package `SnowballC`.

Text placeholders The raw text data had a large number of non-word tokens such as numbers, dates and URLs. Most unique numerals and URLs did not appear more than once in the data and some numerals such as years do not contain information without a context. For this reason, numerals and URLs were replaced by placeholders listed in Table 3.4.2.

Table 3.4.2: Placeholders of commonly occuring numerals in discussion forum text data that were replaced during data preprocessing.

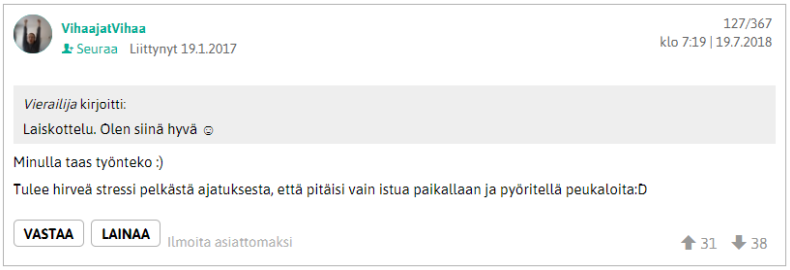
Title	Placeholder name	Replacement rule
Year	A word (starting the string or after whitespce) starting with digits "19" or "20" followed by exactly 2 digits and ending in whitespace or end of a string.	#year
Date	A word with three digits separated by dots.	#date
Numeral	If not matched to previous items in this table: a word of one or more digits, one or more digits followd by a single punctuation and then one or more digits, one or more digits followed by either one or more alphabetical characters or punctuation and one or more digits and then one or more alphabetical characters.	#numeral
URL	Word starting with either *http* or *www*. Not successful for all URL entries.	#URL

Emojis *Emoticons* are facial expressions represented by characters in text used to help the author to express emotions and feelings such as :) and :-(. *Emojis* are

much like emoticons but they are actual small images representing not only facial expressions but a wide range of other objects such as persons, objects, animals, flags, signs etc. In addition to emojis provided by specific mobile apps and devices, Unicode encoding has support for a wide range of emojis available for use on platforms supporting Unicode (Davis and Edberg, 2018).

Discussion forum posts included a notable amount of emojis in Unicode format and some emoticons as illustrated by Figures 3.4.1a and 3.4.1b, respectively. Due to the large number of different varieties of emojis, they were replaced by placeholder values representing the sentiment and basic type of the emoji: face or other symbol. Sentiment and type mapping was based on work by Kralj Novak et al. (2015) where the sentiment of each emoji was modelled by a 3-simplex over "positive", "negative" and "neutral" using a dump of Twitter posts as source data. To obtain a suitable small number of different emoji placeholders, each emoji was mapped to the sentiment which had the largest value on the simplex. Examples of the mapping are found in Table 3.4.3. Emoticons were first mapped to their emoji counterpart and then to sentiment. However, only the most usual emoticons were included in this application. If multiple emojis were typed sequentially in a post, they were replaced by individual placeholder values for each emoji in the sequence.

(a)



(b)

Figure 3.4.1: Examples of messages on `vauva.fi` with emojis and emoticons. The first example features the use of the most popular emoji of the dataset: "laughing with tears" and the second one features the use of two popular emoticons ":)" and ":D" which represent a smiling and a laughing face. Both examples are screen captures from the website accessed on 25.7.2018.

Stop words Stop words are frequently occurring words with little meaning such as "and", "if" or "always". They are removed from the corpus in order to make processing faster. The stop word list is language and application specific. In this application where topics with semantic meaning are studied, a long list of adverb, pronoun

Table 3.4.3: Example of mapping emojis to broader placeholder terms. Sentiment of each emoji is derived from (Kralj Novak et al., 2015). Emojis were separated to face symbols and other symbols and to three sentiment categories: positive, neutral and negative.

Emoji	Sentiment	Category
	Positive	Face-like
	Positive	Face-like
	Negative	Face-like
	Negative	Face-like
	Neutral	Symbol
	Negative	Symbol

and particle words were removed. A Finnish stop word list from GitHub repository *Stopwords-ISO* containing 847 words was used.

Vocabulary pruning At the end of the text preprocessing, the vocabulary is *pruned*. Terms that are too often or too rarely occurring are removed from the vocabulary entirely. Too frequently occurring terms indicate that the word conveys little semantic information that is relevant in the context of the corpus. Same goes for too infrequently occurring terms: they are noise in the corpus, providing little information. Pruning rules used in this thesis were pruning terms that occur in more than 20 % of the documents of the corpus and terms that occur less than in 50 documents in the corpus. This resulted in a vocabulary of 48 594 terms for `suomi24.fi` and 42 905 terms for `vauva.fi`.

3.5 Topic modelling and trend identification

Two topic models were used in this thesis application: a Latent Dirichlet Allocation topic model and a Dynamic Topic Model. After fitting the models on the Internet discussion forum data, trend events were identified from the proportional topic volume time series using a method developed for this thesis. Software solutions for the topic model fitting, the method for parameter tuning for the models as well as the trend identification method are presented in this subsection.

3.5.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation topic model was selected as the primary topic model for the application. There are many factors that favour LDA against other topic models. It is the most widely used topic model in the literature and has comprehensive open source software implementations readily available. LDA may not capture as much detail and information about the data as some other topic models do (cf. the Correlated Topic Model, which models also topical correlations (Blei and Lafferty, 2005)) but the difference has been shown to be small, when using human evaluation (Chang et al., 2009). The background of the LDA model is described in Subsection 2.2.

The software implementation of LDA used in this thesis was the `text2vec` -package of `R`, which implements the fast *WarpLDA* inference method (Selivanov, 2016). The package includes also fast vectorized vocabulary and token handling, which was used in the preprocessing. `text2vec` -package was primarily selected, because of the fast *WarpLDA* implementation. The data size was large at roughly 2 million documents for both data sets and final vocabulary size of over 40 000 terms. The vanilla LDA inference using e.g. collapsed Gibb's sampling is implemented in other packages, such as the `lda` -package, but their performance was substantially slower. The fast LDA implementation was useful, when tuning parameters and exploring different topic numbers. Moreover, the `text2vec` -package included a connector to the `LDAvis` -package and visualization tool previously presented in this thesis (Sievert and Shirley, 2014). The *LDAvis* -tool was useful in early topic exploration and validation work.

3.5.2 Dynamic Topic Model

The DTM was the second topic model used in the thesis application. DTM extends the LDA in such way that the topic mixtures are allowed to vary over time (Blei and Lafferty, 2006). This allows the modelling of dynamic development of topic contents between documents divided in time slices based on their occurrence. The development and variations in the topic mixtures are very important from the viewpoint of assessing, how viable it is to use a topic model that does not allow changes in the topic mixtures (such as the LDA) on data that covers a long period of time. The model is covered in Subsection 2.2. The original paper by Blei and Lafferty (2006) presented two inference methods based on variational Kalman filtering and variational wavelet regression. The software implementation used in this thesis

was developed by the original authors and is available in the `blei-lab` GitHub repository. The implementation is written in C and the inference uses variational wavelet regression.

The data was preprocessed in a similar way as for the LDA in `R`, but then exported to be used in the DTM C-code. Inferencing the dynamic topic model took significantly more time than the LDA because no memory optimized, fast implementations similar to WarpLDA were available. The results were written in files and then read again into `R` for analysis and visualization. Because of differing result data formats, the results of the DTM could not be visualized with the *LDAvis* -tool.

3.5.3 Model evaluation and selection

Data preprocessing parameters were experimented with, but no structured tests were conducted. Evaluating the effects of different data preprocessing parameters in a structured manner would have required complete LDA results to be calculated and evaluated. This made the testing prohibitively time consuming and only qualitative tests were conducted.

LDA models were evaluated both by model training data and a held out test data perplexity as well as using human evaluation. Model evaluation was done with respect to the number of topics in the model and the model convergence tolerance. The final LDA models were selected based on a structured test scheme.

DTM models were fitted and evaluated after LDA models and based on the results obtained with the LDA models. The main study was about the viability and necessity of using the more complex DTM instead of the LDA topic model. This was studied by qualitative evaluation of the modelling results. The method and the results are presented in Subsection 4.1.3.

LDA Model convergence tolerance Fitting the LDA model with WarpLDA implementation is very fast compared to plain EM and Gibbs sampling inference used in basic implementations, but it still took tens of minutes to fit a model to `vauva.fi` or `suomi24.fi` discussion data. A critical parameter affecting model fitting time was the convergence tolerance, i.e. the proportional change in model log likelihood allowed between the iterations. In order to find a suitable convergence tolerance parameter, a test was conducted, where the model perplexity was measured for a number of different convergence tolerances. The topic number was kept constant at 20. The tested convergence tolerance values were $t = \{0.1, 0.01, 0.001, 0.0001, 0.00001\}$. The held out test data consisted of a 10% random sample of the entire data set. The test was conducted only with `vauva.fi` data between 1.5.2018 and 30.11.2018.

It has been noted before in this thesis that the model perplexity (likelihood) has been shown to be an unreliable metric for topic model performance and even to have a negative correlation with model interpretability (Chang et al., 2009). Despite this, perplexity is still used as a metric for tuning the model parameters, because human evaluation is very resource intensive and slow. In the case of convergence tolerance, perplexity was seen as a sufficient metric to look at, when tuning the parameter. For

the number of topics, which has greater effect on model interpretability and usability, also human evaluation was used besides perplexity.

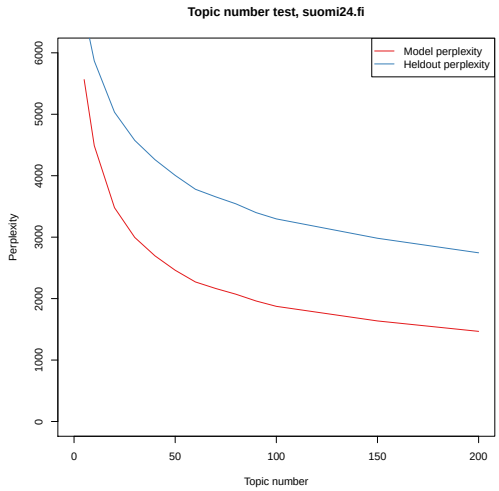
LDA Model topic number evaluation Topic number has a great effect on the interpretability of the model. A low number of topics might produce too general topics that do not adequately describe the documents and too many topics might lead to a high number of topics that do not make sense. This is analogous to the bias-variance tradeoff encountered in machine learning. A test was conducted to find a suitable number of topics for the application. Models for both `vauva.fi` and `suomi24.fi` were fitted with various topic numbers. The models were assessed both from the viewpoint of trained model and held out data perplexity as well as a human evaluation of topics.

The human evaluation was loosely based on the evaluation experiment by Chang et al. (2009), where *word intrusion* and *topic intrusion* were defined. Due to the lack of resources, the only human evaluator in this thesis was the author.

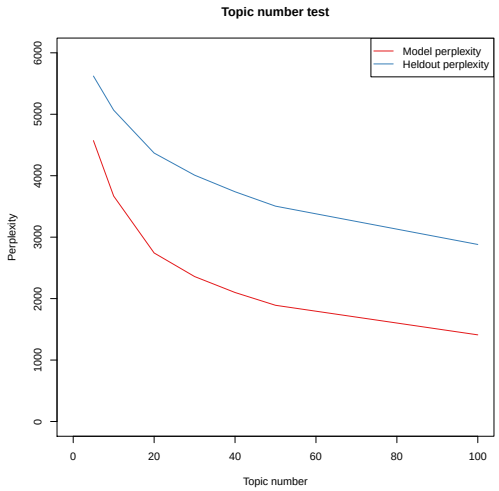
Word intrusion in the original experiment by Chang et al. (2009) was tested by checking, whether an evaluator was able to distinguish a word with low topic probability, inserted into a set of high topic probability words. Word intrusion measures how well the top words of a topic represent a semantically meaningful concept. In this thesis, word intrusion test was replaced by a simpler test. The top 10 words of a topic, ranked by *relevance* defined by Sievert and Shirley (2014) with $\lambda = 0.6$, were evaluated. If the top 10 words could be easily seen as representing a semantically coherent concept, the topic would "pass" and be given a label. The proportion of "passed" topics for each model with different topic numbers was recorded. This metric was called the *word pass rate*.

Topic intrusion in the experiment by Chang et al. (2009) was measured by checking if a human evaluator was able to distinguish a document with low topic probability, inserted into a set of high topic probability documents, based on the title and the first few sentences of the document. Topic intrusion measures the quality of the document-topic assignments of the model. In this thesis, the topic intrusion test was replaced by a simpler test. A human evaluator (the author) was presented with a sample of three high topic probability documents for each topic that had "passed" the modified word intrusion test. The topic was represented by the label given in the modified word intrusion test. If the document was clearly characterised by the topic label, the document would pass. The proportion of passes for each topic in the model was recorded and the mean of passes for the entire model was also recorded. This mean was called *document pass rate* and it was calculated only over the topics that passed the modified word intrusion test.

Selecting the topic number for the final results was based on both the perplexity of the models and the mean of the topic pass rate and the document pass rate. Perplexity was calculated for models fitted with topic numbers $k \in \{5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200\}$. The results for `suomi24.fi` are in Figure 3.5.1a and the results for `vauva.fi` are in Figure 3.5.1b. It can be seen that perplexity does not decrease significantly after 70 or 50 topics (`suomi24.fi` and `vauva.fi` respectively). The perplexity for under 40 or 20 topics is considerably higher. That



(a) Results for `suomi24.fi`. Increasing topic number decreases perplexity even after 150 topics. However, the decrease in perplexity is small after ca. 70 topics.



(b) Results for `vauva.fi`. Increasing topic number decreases perplexity even after 50 topics. However, the decrease in perplexity is small after ca. 50 topics.

Figure 3.5.1: Perplexity for fitted LDA model and held out documents with varying number of topics. The data used for the model convergence test consisted of all discussion posts from 1.5.2018 to 30.11.2018. The test set comprised 10% of held out data.

is why the human evaluation was completed and topic and document pass rates were calculated only for models with $k \in \{40, 50, 60, 70\}$ for `suomi24.fi` and $k \in \{20, 30, 40, 50\}$ for `vauva.fi`. Completing the evaluation for both data sources, with the aforementioned topic numbers, required the evaluation of 360 topics and 1080 documents by hand, which was a substantial task. However, the utility was twofold: In addition to the selection of the most suitable topic number, the assessment of topic quality also provided a list of interpretable topics and topic labels. Subsequently a list of topics with little interest due to semantic incoherence (did not pass with respect to topic pass rate or scored low in document pass rate) was obtained and these "bad" topics could then be excluded from further analysis.

The final metric used for comparing the models with different topic number was the mean of the topic pass rate and the document pass rate. This way both the number of interpretable topics and the quality of topics relatedness to a document were affecting the results.

3.5.4 Trend identification

The topics of the discussion forum or any text data can be of much interest on their own. However, it is even more interesting to know *when* a topic is especially important and why. Thus, a method for identifying the topical *trend events* of the discussion forum data was developed for this thesis. A trend in this case is not a long time trajectory of growth or decline, but rather a sudden increase of topical interest.

The hypothesis is that these trend events can be identified by a clear increase from typical deviations of proportional topic volume in the data. For this, the proportional topic volumes are calculated for each date in the discussion data, such that the topic mixtures of documents for a given date are summed from the topic model fit. Proportional topic volume is normalized over the daily discussion post volume.

Trend events The solution for trend event detection was developed to detect a period of time where the topic volume for some topic is abnormally large. The method resembles outlier detection. The principle idea of the method is given below:

- For each topic $z_k, k = 1, 2, ..., K$:
  - Calculate rolling median and rolling median absolute deviation (MAD) of proportional topic volume with lag L. Denote rolling median for date t with $\mathrm{M}_{roll}(t)$ and rolling MAD with $\mathrm{MAD}_{roll}(t)$.
  - For each date t that is *not* already marked as part of a trend event, list a series of time windows with different lengths:

$$T_l = \{t + l_{min}, ..., t + l_{max}\}, l \in \{l_{min} \leq l \leq l_{max}, l \in \mathbb{N}\} \tag{1}$$

  - Select the maximal window length $\hat{l}$, for which the median of the proportional topic volume within the window T_l is be larger than the sum of the rolling median and the MAD multiplied by a threshold coefficient τ. Select the maximal window length such that the condition is met by all l less than the maximal $\hat{l}$. The selection criteria is given in equation (2)

$$\arg\max_{\hat{l}} \ \mathrm{M}(z_{k,T_l}) > \mathrm{M}_{roll}(t) + \tau \mathrm{MAD}_{roll}(t) \ \forall \ T_l, \ l \leq \hat{l} \tag{2}$$

  - The time window that meets the criteria is considered a trend event.

The identification method resembles traditional outlier methods, based on selecting a certain quantile of ordered data values as the threshold for considering a data point an outlier. Since the topic volume data is a time series, a rolling value must be used to determine the outlier threshold. No assumptions on the distribution of the proportional topic volumes can be made, which requires the usage of distribution free location and scatter metrics: the median and median absolute deviation. A trend event occurs only if there is a high volume of conversations as opposed to low or no conversation. This is why the outliers are only searched by looking at abnormally high, not low, topic volumes. Proportional topic volume was used as the time series data, because the data has large absolute variations due to errors in data crawling and e.g. holidays, which affect the visitor volumes on the websites.

It is intuitive to hypothesize that the length of the trend event should not be restricted to one day (which is the time step of the time serialized topic volume data), but longer trend events should be allowed as well. This is why the method has the step of finding the maximal time window, where the threshold criteria is fulfilled for the median of the entire window. This favours finding longer time periods of increased topical activity, which can then be analysed as a single event.

Trend event terms The lag L, the maximal and minimal trend event window lengths l_{max}, l_{min} and the threshold coefficient τ must be tuned to obtain realistic trend events. An evaluation criteria had to be devised. The hypothesis is that if the trend event is of interest, the discussion data contents should reflect the cause of the increased topic volume. The important terms of the trend event should be both especially important for this period of time compared to the rest of the discussions and important in the specific topic under study.

To find out the relevant terms *for a single trend event* within a topic, metrics called *trend lift* and *trend relevance* were developed.

- Trend lift TL: The ratio of the TF-IDF (term frequency - inverse document frequency) values for each term for the trend event and the entire corpus:

$$TL_{t,event} = \frac{TF - IDF_{t,event}}{TF - IDF_t} \tag{3}$$

  - TF-IDF is defined as $TF - IDF_t = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \cdot \log \frac{N}{n_t}$, where $f_{t,d}$ is the frequency of term t in document d, N is the number of documents in the corpus and n_t is the number of documents that contain term t.
  - An alternative for the TF-IDF ratio is to use the ratio of proportional term frequencies for the trend event and the entire corpus. Term frequency is defined as: $\frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$.

- Trend relevance TR: The ratio of the term relevance $r_t = \lambda \cdot log(p_t) + (1 - \lambda) \cdot log(\frac{p_t}{lift_t})$ (applies for the entire model) and the trend lift raised to the power of the coefficient of the trend lift α:

$$TR_t = r_t / TL_{t,event}^{\alpha} \tag{4}$$

The trend relevance TR_t is close to zero for terms that have both high relevance r_t, meaning that they are important terms in the topic, and high trend lift TL_t, meaning that they have pronounced importance during the trend event in question. Relevance itself is a negative number with high relevance corresponding to values close to zero, which explains the use of division instead of multiplication for calculating the TR. The descriptive words for the trend event are then selected as the top words ranked by trend relevance. These words can then be used to tune trend identification parameters as well as to interpret the meaning and implications of the trend events.

The trend relevance metric can be small even though the absolute number of term occurrences during the trend event is low. For this reason, an additional criteria of minimal number of documents during the trend event containing the term was included. This was sensible, as some high ranking trend event terms had only few document occurrences, meaning that they can not be part of a more widespread phenomenon. The coefficient (exponent) α of the trend lift TL_t metric had to be included in the model in order to tune how much the topical relevance and the temporal distinctiveness of the term affect the final trend relevance.

Trend event visualization The trend events must be visualized in such a way that the user may easily detect the time and length of the event and see the events' descriptive trend terms. Since there are multiple topics and multiple trend events for each topic over time, it is necessary to be able to present multiple trend events at the same time, but not too many to confuse the user.

The graph to visualize the results was selected to be based on the original proportional topic volume time series. Then the rolling median and the sum of the rolling median and the rolling MAD multiplied by the threshold coefficient are plotted to indicate the basis for the trend event detection. The trend events are marked by colouring the parts of the proportional topic volume line graph that belong to a trend event red. The top 8 trend event terms are displayed in the graph above the trend event. An example of the trend event visualization for a topic in the `vauva.fi` -model is in Figure 4.2.2.

Trend identification parameter tuning The trend identification method included four parameters that affect the results: the lag L, maximal and minimal trend window length l_{min} and l_{max} and the threshold coefficient τ. The parameters were tuned to have a reasonable amount of trend events as well as to find an understandable connection between the trend event terms and real world events. Lags $L \in \{7, 15, 21, 28\}$, minimal trend window lengths $l_{min} \in \{1, 2, 3\}$, maximal trend window lengths $l_{max} \in \{4, 5, 6, 7, 8\}$ and threshold coefficient values $\tau \in \{1, 1.5, 2, 2.1, 2.5\}$ were tested such that the values of a single parameter were varied while others were kept constant at a default setting $L = 21$, $l_{min} = 2$, $l_{max} = 6$, $\tau = 2$.

A single parameter related to the selection of trend event terms, the coefficient of the trend lift α, was also tuned. It is the exponent of the trend lift metric developed to indicate how specific a term is to a given period of time. Increasing the coefficient of the trend lift results in trend event terms that are more specific to the time period in question rather than the topical relevance.

The number of trend events, number of trend events with little time in between, the coherence of the trend word list and the connection of the trend word list to known real world events were recorded. A qualitative assessment of these factors was conducted to determine suitable parameters for the trend identification method.

4 Topics and trends identified in Finnish Internet discussion forums

An LDA topic model as well as a Dynamic Topic Model were fitted to discussion data of both `suomi24.fi` and `vauva.fi` Internet forums from a time period of 1.5.2018 to 30.11.2018. A series of fits with different topic numbers were calculated and the topics evaluated. After tuning the models with respect to topic number, topical trends were identified and each trend event was associated with trend words describing the event, based on the trend relevance metric. Trend identification effectiveness was evaluated by comparing the discovered trend events to real world events. It should be noted that all of the terms presented in the results are *stemmed* (as stated in the data preprocessing pipeline), which explains why they might not directly correspond to Finnish words, but rather to shortened versions of them.

Experimental setup `suomi24.fi` and `vauva.fi` data were handled separately and the experimental setup was replicated for both data. The data size for both forums is stated in Table 4.0.1.

Table 4.0.1: Data sizes for `suomi24.fi` and `vauva.fi` topical trend identification application. The data time period is from 1.5.2018 to 30.11.2018.

Forum	Posts	Vocabulary
`suomi24.fi`	1 656 823	48 594
`vauva.fi`	2 438 593	42 905

- LDA fitted with WarpLDA, convergence tolerance tuned as described in 3.5.3
- LDA fitted with WarpLDA, topic number tuned as described in 3.5.3 (perplexity and human evaluation)
- LDA model topics evaluated and studied
- DTM, single model fit and study
- LDA models, trend event threshold tuned with qualitative human evaluation
- LDA models, trend events evaluated and studied

Computational platform Data preprocessing as well as LDA model fitting and evaluation were done on a Google Cloud virtual machine with RStudio `R` development environment. DTM fitting was done on a laptop PC with Windows Subsystem for Linux running Linux Ubuntu. Computer specifications for both platforms are listed in Table 4.0.2.

Table 4.0.2: Computational platforms used in the application

	Google Cloud VM	Laptop PC
Used for	Preprocessing, LDA fitting	DTM fitting
Operating System	Debian Linux	Ubuntu Linux
Logical CPU cores	8	4
CPU type	Intel Xeon E5 v3	Intel Core i5-7200U
CPU frequency	2.3 to 2.8 Ghz	2.5 to 3.1 Ghz
System memory	52 GB	16 GB

4.1 Topic modelling results

4.1.1 LDA parameter tuning

The convergence tolerance of the WarpLDA model fit as well as the topic number of the model were tuned according to the test setup described in Subsection 3.5.3.

Model convergence tolerance It was observed that convergence tolerance below **0.001** had no effect on model perplexity performance for either the fitted model or the evaluated model from held out data. This convergence tolerance value was selected for all further WarpLDA model fits. The result is seen in Figure 4.1.1. The test ran a total of 64 minutes on the Google Cloud VM using multiple cores. It was observed that because the WarpLDA implementation of the package `text2vec` was not multi threaded, the test length was dependent on the single core performance of the longest duration test instance (in this case the lowest convergence tolerance). The term number and data types are similar in both `vauva.fi` and `suomi24.fi`, so it was deemed unnecessary to tune the convergence tolerance parameter separately for both data and the value 0.001 was used for them both for all further tests.

Model topic number evaluation Model perplexity for models with various topic numbers are displayed in Figure 3.5.1 for both `suomi24.fi` and `vauva.fi`. The perplexity is significantly higher for models with under 40 or 20 topics respectively. On the other hand, the perplexity decreases slower after increasing the topic number to over 70 or 50. For this reason, as was stated in Subsection 3.5.3, the human evaluation was performed only for topic numbers $k \in \{40, 50, 60, 70\}$ in `suomi24.fi` and $k \in \{20, 30, 40, 50\}$ in `vauva.fi`. The results of the human evaluation (as described in Subsection 3.5.3) are presented in Figures 4.1.2 and 4.1.3. The mean of the topic pass rate and the document pass rate was used as the model selection criteria. It turns out that 60 topics in `suomi24.fi` and 40 topics in `vauva.fi` provide the most human interpretable results. The higher topic number in `suomi24.fi` indicates that the forum contains more varied selection of topics compared to `vauva.fi`.

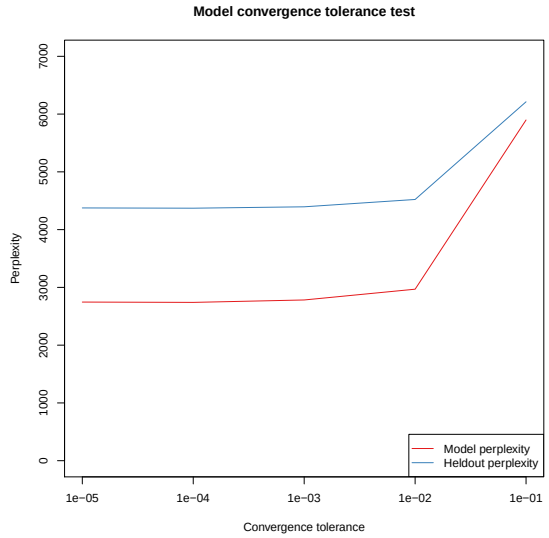


Figure 4.1.1: Perplexity for fitted LDA model and held out documents with different model fitting and evaluation convergence tolerances. X-axis is in log scale. Even with the log scale, a slight elbow point can be identified at either 0.01 or 0.001. The data used for the model convergence test consisted of all discussion posts from `vauva.fi` from 1.5.2018 to 30.11.2018. The test set comprised 10% of posts held out of the entire data.

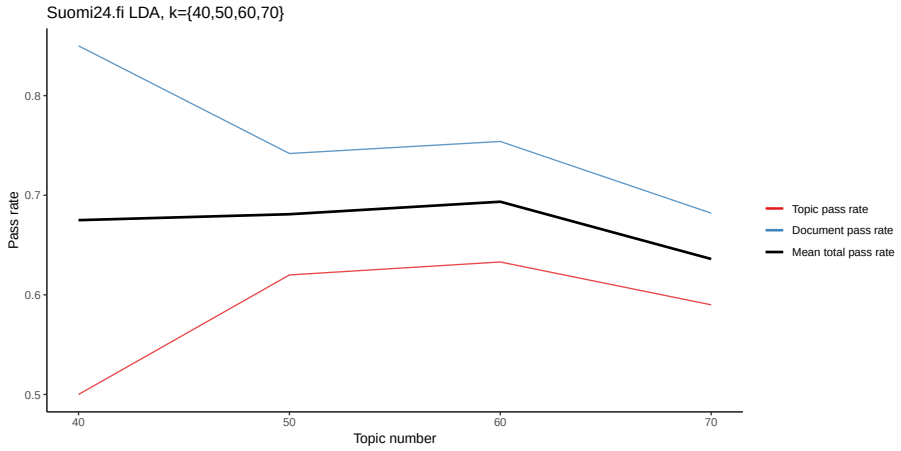


Figure 4.1.2: The topic pass rate and the document pass rate for `suomi24.fi`. The mean of the two rates was used the criteria for topic number selection. Maximal mean rate was achieved with 60 topics.

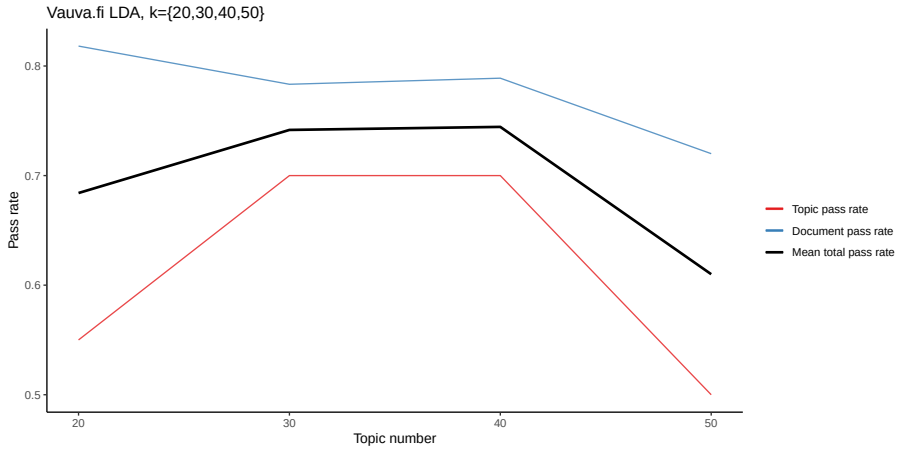


Figure 4.1.3: The topic pass rate and the document pass rate for `vauva.fi`. The mean of the two rates was used the criteria for topic number selection. Maximal mean rate was achieved with 40 topics.

4.1.2 LDA topic interpretation

The topics obtained by LDA topic models fitted in `suomi24.fi` and `vauva.fi` data are presented in this subsection. While `suomi24.fi` has been presented first in the previous section of this thesis, the `vauva.fi` results were more interesting, both in LDA topic results and topical trend identification (see Subsection 4.2). This is why more emphasis is put on `vauva.fi` LDA topic and topical trend results, and they are presented before `suomi24.fi`.

Topics in `vauva.fi` The final LDA model for `vauva.fi` data decomposed the forum posts into 40 topics. Of the 40 topics, 30 (75%) topics were interpreted to have a coherent semantic meaning, based on the 10 top words given by the *relevance* metric with weighting factor $\lambda = 0.6$. These topics were given a short label. All 40 topics are listed in Figure 4.1.4. Table 4.1.1 includes English translation for the topic 2, "Food", as an example. The rest of the topics are not translated, but an English label is given instead. The small black circle represents the summed proportion of a topic in all documents. The largest 5 topics with a label are listed in Figure 4.1.5. It should be noted that while the top 5 largest topics represent different themes, the summed proportion of all topics related somehow to relationships comprise a much larger proportion of the topic space than any of the top 5 largest topics. This is illustrated in Table 4.1.2. This indicates that much of the discussion on `vauva.fi` is related to the daily struggles in families and relationships, as is the intended purpose of the forum itself. The top words by relevance in the topics contain mostly very general words and only few names or specific nouns. To improve the detection of more meaningful terms, more emphasis should be put on vocabulary pruning, selection of the term importance metric and source data preprocessing (e.g. lemmatization and removal of verb words). In the data preprocessing, emoticons and emojis were replaces by placeholder values to decrease the number of unique terms. It is notable, the the placeholder terms occur in only a couple topics' top word lists. Thus, it seems that emojis and emoticons are used in a limited number of posts and typically they comprise a comparatively large proportion of the post they occur in. More suggestions for further improvements are discussed in Subsection 5.1.

Table 4.1.1: English translations for the 10 top words in `vauva.fi` topic 2, "Food". The shortcomings of stemming are evident, as many of the terms represent the same lemma.

Original terms	Translation
ruoka	food
syödä	to eat
syö	eat
kahv	coffee
liha	meat
syön	I eat
syömä	to eat
ruokavalio	diet
herku	delicacy / to feast
nälk	hunger

Table 4.1.2: The proportional topic volume of top 5 largest topics, which had a coherent meaning, and the sum of the volume of all topics related to realtionships and family in `vauva.fi`. Even though none of the relationship themed topics are among the top 5 largest topics, they amount to 15% of the total topical volume in the discussion. This reflects well with the nature of the `vauva.fi` discussion forum.

Topic	Volume
Relationship topics total	0.141
"Politics"	0.030
"Media"	0.028
"Food"	0.027
"Social media"	0.027
"Looks"	0.026

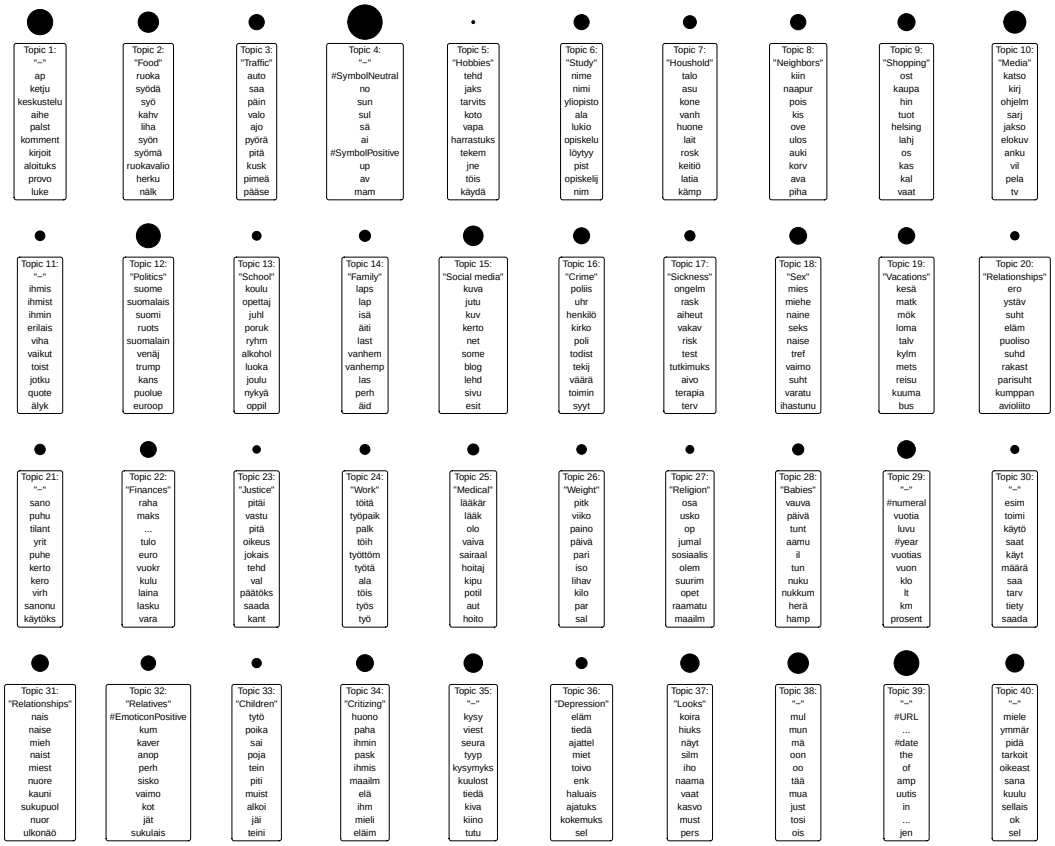


Figure 4.1.4: All 40 topics of the LDA topic model fitted in the `vauva.fi` data. The top 10 words by relevance are listed and the relative size of the topic is displayed by the black circle.

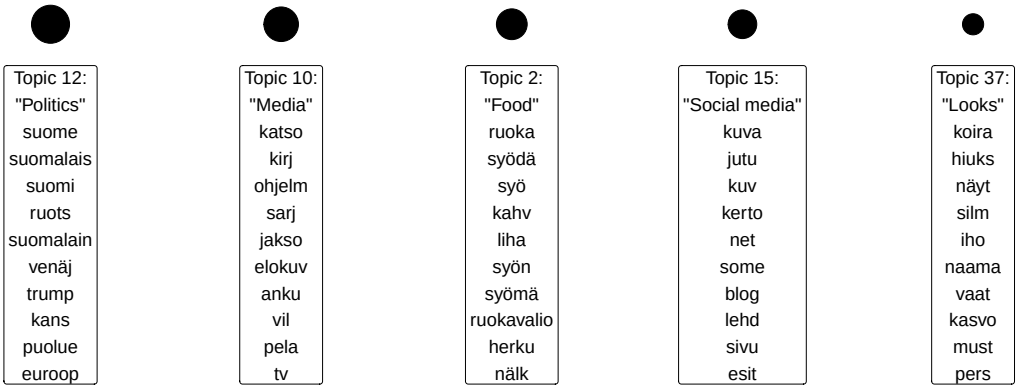


Figure 4.1.5: Top 5 largest topics of the 40 topic LDA model of `vauva.fi` that were interpreted to have a coherent meaning.

Topics in `suomi24.fi` The final LDA model for `suomi24.fi` data decomposed the forum posts into 60 topics. Of the 60 topics, 37 (62%) topics were interpreted to have a coherent semantic meaning, based on the 10 top words given by the *relevance* metric with weighting factor $\lambda = 0.6$. These topics were given a short label. The proportion of labelled topics was lower than in `vauva.fi`, which had 75% coherent topic proportion. All 60 topics are listed in Figure 4.1.6. The largest 10 topics with a label are listed in Figure 4.1.7. The top 10 topics represent quite different themes and indicate that the discussion in `suomi24.fi` covers a broad spectrum of topics. In comparison to `vauva.fi`, not as many of these topics show as much potential for business or marketing use. For example, topics about food and media are missing in the top 10 topic list.

By looking at the entire topic space, it must be noted that some themes occur repeatedly with only a little difference in the topics' top word selection. Topics about politics and religion seem to be especially common in the `suomi24.fi` discussion. Table 4.1.3 summarizes the total proportional volumes of all politics and religion related topics in `suomi24.fi` as well as the top 5 topics as a point for comparison.

The topic model included some very specific topics, which could potentially have more use in marketing and business For example, the topic about cars (topic 56) and personal electornics (topic 40) are both viable for relevant signals about the customers' opinions.

The overall tone in politics and religion topic top word lists exhibits obscene language, prejudices and even racism. This seems to be a common feature observed in `suomi24.fi`. In addition to the contents of labelled topics, there are a lot of topics without a clear coherent theme but rather they describe a portion of meta talk on the forum about what posts should be discarded or reported to the administration (e.g. topic 11) or offensive comments about other people (e.g. topic 19, which contains mostly swear words in the top word list). While similar content was observed in some of the `vauva.fi` topics, the phenomenon seems to be more common in `suomi24.fi`.

Table 4.1.3

Topic	Volume
Religion topics total	0.086
Politics topics total	0.108
Sex	0.021
Domestic Politics 2	0.019
Travel	0.019
Pets and Neighbors	0.018
Seasons	0.018

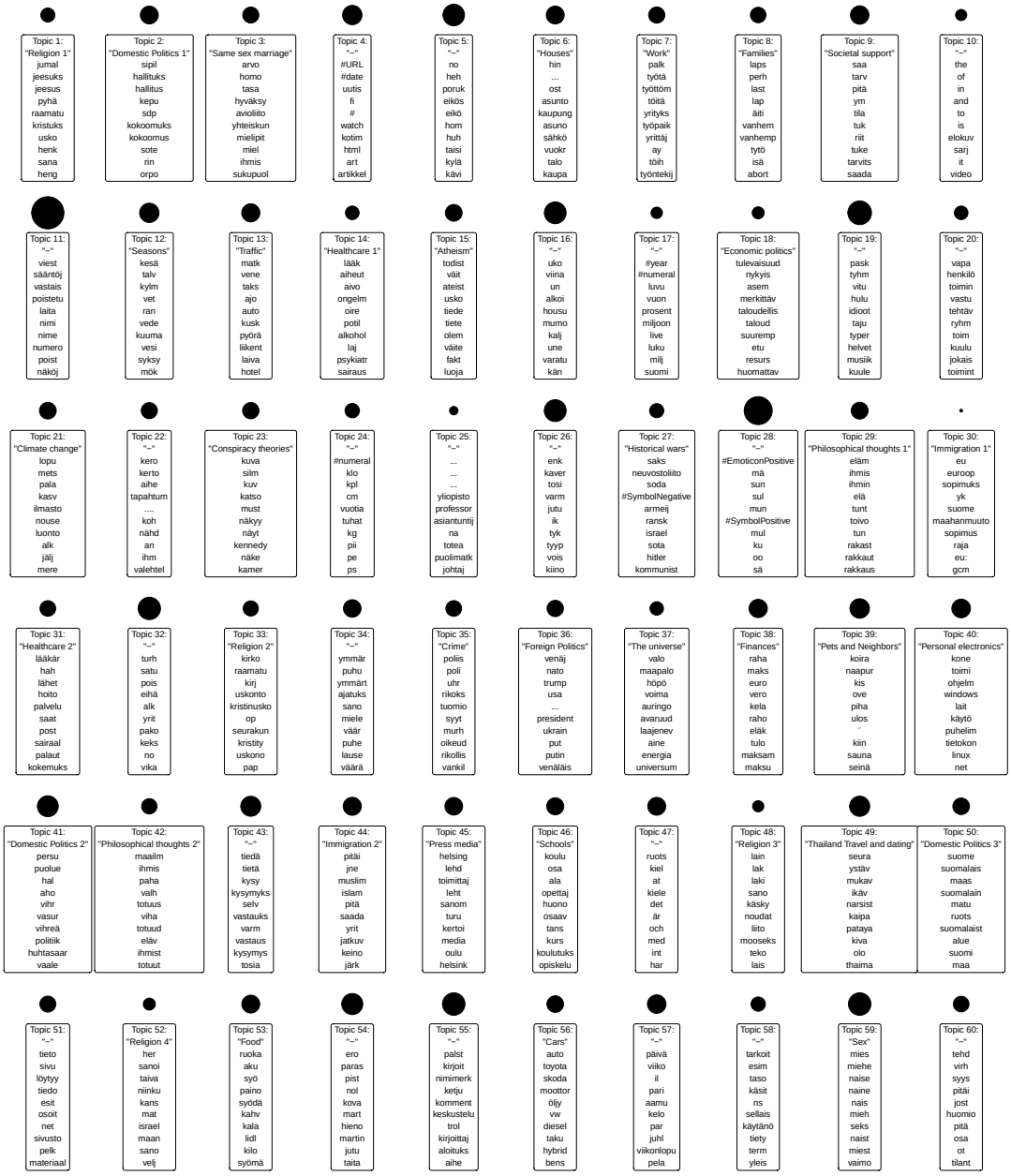


Figure 4.1.6: All 60 topics of the LDA topic model fitted in the `suomi24.fi` data. The top 10 words by relevance are listed and the relative size of the topic is displayed by the black circle.

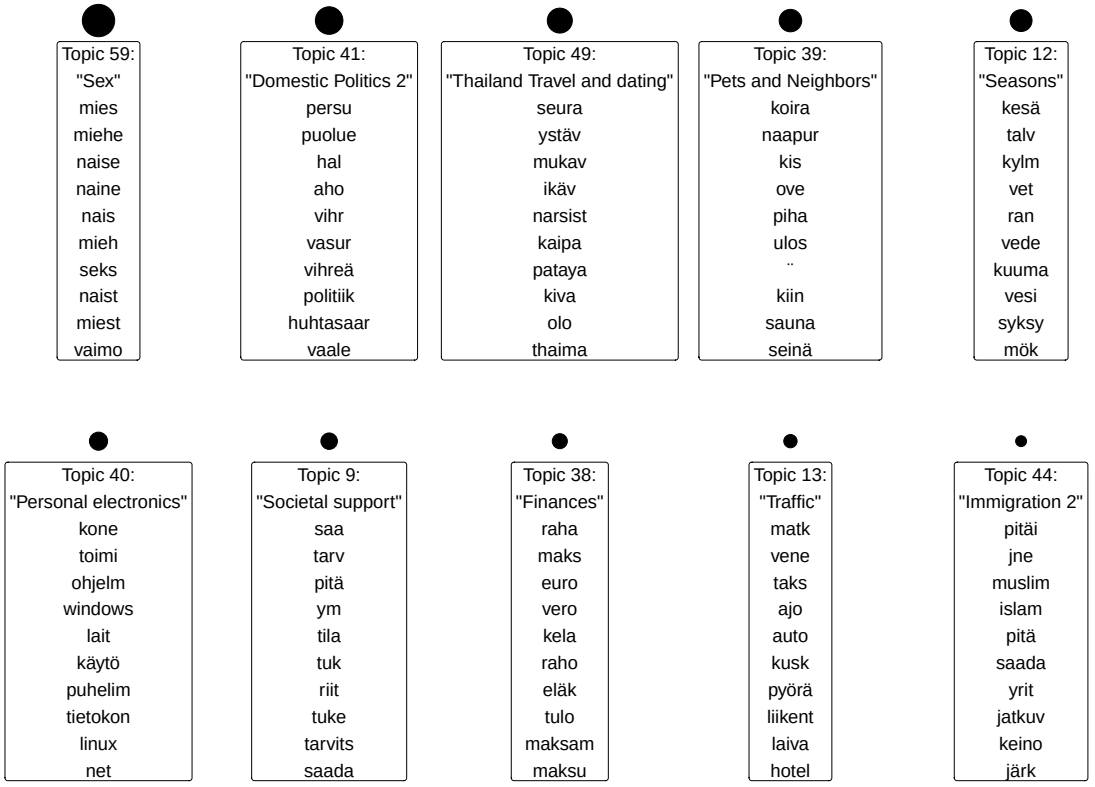


Figure 4.1.7: Top 10 largest topics of the 60 topic LDA model of `suomi24.fi` that were interpreted to have a coherent meaning.

4.1.3 Dynamic Topic Model results

Dynamic Topic Model (DTM) was also considered as a topic model for decomposing the text data into topics and consequently to identify the topical trends. A DTM implementation written in C by David Blei is published in the `blei-lab` Git repository and it was used for fitting the model. The data for DTM tests was a sample of 500 000 posts from both `vauva.fi` and `suomi24.fi` from Summer 2018, spanning 11 weeks from May to July. The vocabulary was pruned to have ca. 17 000 terms. The data size was limited due to the computational intensity of the DTM fitting. The topic number used for the test was 30 due to the findings in `vauva.fi` LDA model fitting, where 30 and 40 topics scored quite close to each other in the human evaluation (see Figure 4.1.3) and a lower topic number was preferred for faster model fitting. The Dynamic Topic Model allows the topic term mixture to vary between discrete time intervals, capturing more information about the development of terms' importance in a topic. The selected time slicing in the test was one week. This resulted in a total of 11 topic mixtures for each topic, one for each week of the used discussion data.

The model fitting time on a PC laptop, with specs given in Table 4.0.2, was over 12 hours. Taking into account that the document quantity was a fifth of the `vauva.fi` LDA topic model and the vocabulary size was roughly 40 % of the LDA topic model vocabulary sizes, the run time of the DTM fitting was unpractically high. For continuous updating and use of the model, a parallel version of the DTM with preferably more efficient posterior inference would be needed.

The added benefit of the DTM over LDA is that if the topics contents vary a lot over time, the DTM assures better fit by allowing the topic mixtures to vary accordingly. However, in the DTM model fit, it was observed that the topic mixtures did not in fact vary much during the time interval of 11 weeks. Example terms from a topic about "Food", with different topic mixture probabilities, are presented in Figure 4.1.8. Observing the top 10 words by topic mixture probability, also indicates that the mixture of the top terms remains very static, through the time slices. It can be seen that the percentual variation in during the time period only ranges from 20 to 50 percent and that the order of term probabilities remains roughly the same. The top words of 11 topic mixtures of each time slice for an example topic are shown in Table 4.1.4. The mixtures show little variation in top words during the time period. These results indicate that for a time period this short, the topic mixtures do not change in such significant quantities that the use of a DTM instead of a LDA topic model would be necessary.

In conclusion, the DTM did not add significant benefit over the LDA topic model and the computational requirements of fitting the DTM were considerably larger than that of the LDA models. Thus, it was concluded that the use of only LDA model fits for the topical trend identification method was sufficient.

Table 4.1.4: Example of topic top word development over time in the Dynamic Topic Model. Each column represents a time slice of one week. The topic mixture does not seem to change much, at least with respect to the most common words in the topic.

	1	2	3	4	5	6	7	8	9	10	11
1	ruoka	ruoka	ruoka	ruoka	ruoka	ruoka	ruoka	ruoka	ruoka	ruoka	ruoka
2	syödä	syödä	syödä	syödä	syödä	syödä	syödä	syödä	syödä	syödä	syödä
3	syö	syö	syö	syö	syö	syö	syö	syö	syö	syö	syö
4	kahv	kahv	kahv	kahv	kahv	kahv	saa	saa	saa	saa	saa
5	saa	saa	saa	saa	saa	saa	kahv	kahv	kahv	liha	liha
6	ost	ost	liha	liha	liha	liha	liha	liha	liha	kahv	kahv
7	liha	liha	sinkku-mies	sinkku-mies	tuot	tuot	tuot	ost	ost	tuot	tuot
8	herku	herku	ost	tuot	sinkku-mies	ost	ost	tuot	tuot	ost	ost

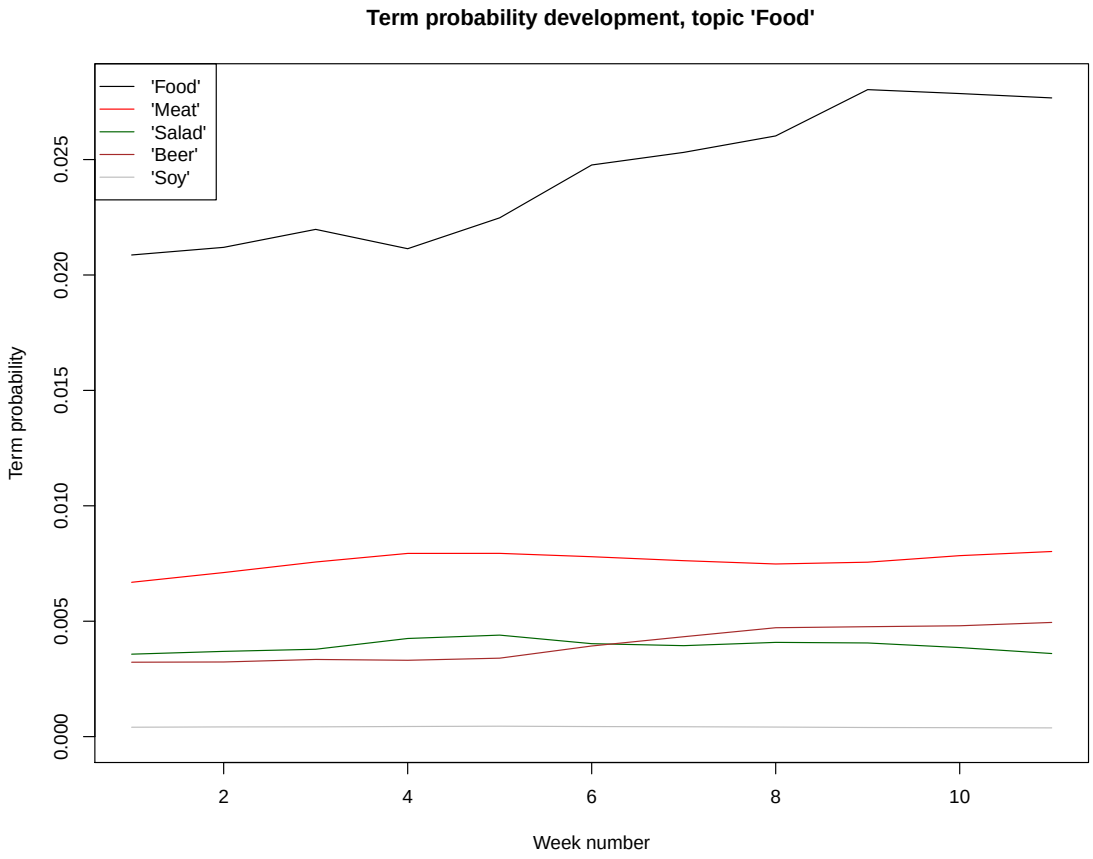


Figure 4.1.8: An example of topic mixture probabilities for a selection of terms in the DTM topic 'food'. The order of the probabilities remains much the same throughout the 11 weeks time span and the variation is restricted.

4.2 Topical trend events

Topical trend events were identified and trend terms calculated for each LDA model topics, which were evaluated to have a coherent meaning. This meant 37 topics for `suomi24.fi` and 30 topics for `vauva.fi`. The identification method as well as parameter selection is described in Subsection 3.5.4. Trend identification parameters were shared with both forums. Main findings and example topics are presented in this subsection. Proportional topic volume time series and identified trend events for all topics, which had a coherent meaning, can be found in Appendices A and B.

4.2.1 Trend identification parameter tuning

Parameter tuning of the trend identification method was studied qualitatively without a formal framework. Tests were conducted on different parameter combinations related to the lag of the rolling median and median absolute deviation L, the trend event threshold coefficient parameter τ (coefficient for rolling mean absolute deviation added to the rolling median to have the threshold for a trend event), the minimal length of the trend event l_{min} and the coefficient of the trend lift in determining trend term α. The main finding was that increasing the minimal length of the trend event resulted in the loss of some important but short events and made the events more general. In addition, increasing the lag of the rolling median and median absolute deviation made the trend events related more to the general variation in topical volume and not local variations. The threshold coefficient parameter mainly affected the amount of the identified trend events and their generality. The exponent parameter of the trend lift had to be increased to obtain more specific trend terms instead of having mostly very general topical terms.

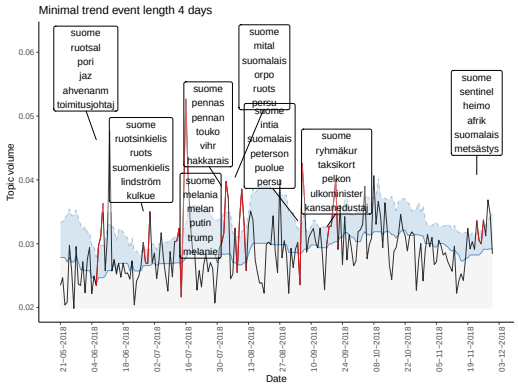
The final value for the minimal length of the trend event was 2 days, since larger values resulted in too general trend events. This means that more events had trend terms that did not add up to any meaningful entity. On the other hand, accepting trend events of one day resulted in an abrupt increase in the number of trend events and more noise in the trend event terms.

The threshold coefficient parameter was tuned to have value 2.1. This resulted in on average 10 trend events per topic during the time period of study (from May 2018 to November 2018). Lower parameter values resulted in an increase of generality in the trend event terms for many trend events.

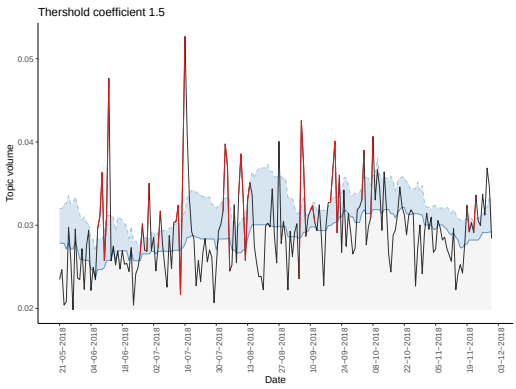
The exponent parameter of the trend lift was decided to be 2. This resulted in trend terms, which differed enough from the top word list of the entire topic. A comparison of trend events for different parameter combinations are displayed in Figures 4.2.1a-4.2.1d , which features the `vauva.fi` topic "Politics" as an example.

4.2.2 Trend event results

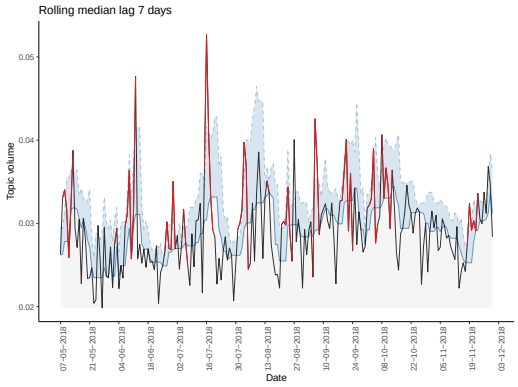
As with the LDA results in Subsection 4.1.2, the topical trend events are first presented for `vauva.fi`, due to their more relevant content, and then for `suomi24.fi`.



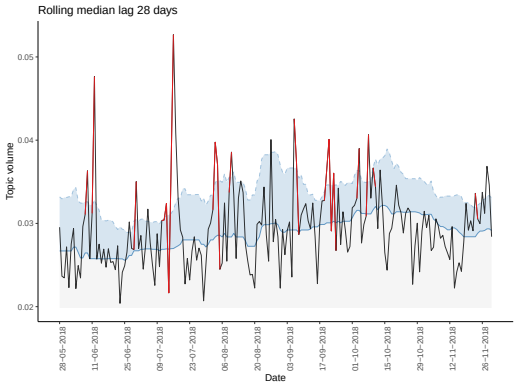
(a) Minimun trend event length set to 4 days. Some important events are left out due to their short duration (e.g. beginning of October, about the greenhouse gasses and the minister for traffic in Finland, see 4.2.6). A shorter minimum length should be selected.



(b) Threshold coefficient set to 1.5. The threshold is quite low, resulting in a high number of short events.



(c) Rolling median and median absolute deviation lag of 7 days. This short value causes the identification method to spot very local increases in the time series.



(d) Rolling median and median absolute deviation lag of 28 days. The long lag creates a higher and more stable threshold for the trend events.

Figure 4.2.1: Examples of the effets of the trend identification method parameters on the identified trend events and trend terms.

Topical trend events in `vauva.fi` Many of the `vauva.fi` topics were found to be related to families, relationships and child care in Subsection 4.1.2. However, topics about other themes were discovered as well. Topics about food, media, personal finances, shopping, vacations and politics could be of use for business and marketing. The main finding from `vauva.fi` trend identification is that for many topics, relevant trend events could be identified. However, not all identified events seem to be relevant, and there are topics that have few or no interesting trend events.

The trend identification method for the forum `vauva.fi` shows great potential. For multiple topics, the method was able to identify trend events and related trend terms that connect to real world events and can be interpreted to contain relevant information. In the topic "Food", displayed in Figure 4.2.2, many real world events can be observed and some trend events exhibit interest in some food ingredients. The time around the graduation day of Finnish schools is the largest trend event and it contains terms about different typical dishes in celebration. The order of the importance of the dishes may be of interest for the food or bakery industry. The discussion about the proposed Finnish army vegetarian meal day is identified (end of August) as is the strike of the Finnish public services union (JHL) and its effect on the school meals (in the middle of October). It is noteworthy that multiple trend events contain terms that relate to vegetarian diets, such as "vegan", "beans", "lentils" and "soy". Towards the end of the study period, two trend events also contain a specific Finnish brand in the food industry, Fazer.

A topic about "Media" included several interesting features, which are seen in Figure 4.2.3. The top terms for many of the the identified trend events are actually artist or celebrity names. In addition, some TV show names also occur, such as "Maajussille morsian" ("Bride for the farmer", event in September) and "Teiniäidit" ("Teenage mothers", event in October). This heightened interest is a useful insight for a company in the TV media business. The topic volume also shows an upward long term trend, which might indicate a stronger overall interest in TV shows and celebrities during the Autumn of 2018. In closer inspection of the trend events during November 2018, it was noted that the trend event terms are actually participant names from the TV reality show "Love Island Suomi". The last trend event co-occurs with the final episode of the show. This is a very interesting insight for the TV media business.

Another example of a topic with business relevant trend events, is the topic about "Shopping". This topic has multiple trend events that contain trend terms about some specific store brands and shopping items. Examples of stores in the trend terms include the clothing store chain*Zara*, web store *Mr. Gugu*, grocery stores *Prisma* and *Lidl* as well as the shopping malls *Redi*, *Jumbo* and *Itis*. The largest trend event in early August features terms about the famous *Moomin mugs*, which were in headlines and discussion due to a rarity item that came to the stores in 9.8.2018. Black Friday shopping event was also recorded in the last trend event. Other topics exhibiting good performance in capturing current events were e.g. "Studying", where the high school graduation and college acceptance notification dates were identified, and "Crime", with identified events about the drug possession of the artist Jari Sillanpää, the spying scandal of Airiston Helmi and the vandalization of the Hietaniemi graveyard

during the Weekend music festival.

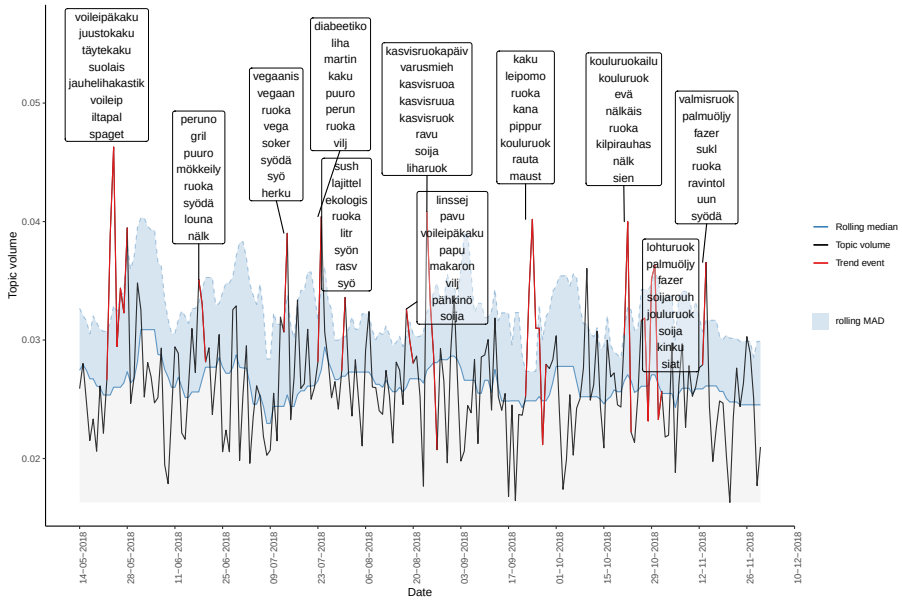


Figure 4.2.2: `vauva.fi` 40 topic LDA model, topic "Food". This topic contains many relevant real world events, such as the school graduation time in May and related dishes, heightened discussion about a specific Finnish brand in the food industry (Fazer), the proposed vegetarian meal day in the Finnish army and the strike of the Finnish public services union, which affected school meals.

Investigation of the trend events complements the topic's overall top word list by revealing more things about the contents of the topic. For example, the topic "Looks" has top words that relate to hair, facial features and clothes. However, the top word list also includes the term "dog", but it is the only pet related term in the list. By looking at the identified trend events in the topic in Figure 4.2.5, four out of fourteen trend events are clearly related to dogs, hunting or horseback riding. This indicates that the original topic "Looks" is actually a mix about looks and animal related hobbies. In some cases, it seems that instead of the underlying topic being a mix between multiple themes, only the trend events contain "rogue" terms from other themes. An example of this is the first trend event in the topic "Politics", where the terms only relate to erectile dysfunction medicine. These terms are not present in any of the other trend events of the topic. The topical volume is high for the rogue trend event, which can be seen in Figure 4.2.6. It can be hypothesised that the terms have very high trend lift, leading to the terms to be identified as trend terms. This balancing between the trend lift (temporal distinctiveness of the terms) and the relevance (topical importance) is an important direction for further development in the model. Having terms, which are too general in the topic do not provide insight about the event, but having terms that do not relate enough to the topic also make the trend events less useful.

Some topics have mostly trend events that do not vary a lot from one another. For example, the topic "Sex" has 13 identified trend events but the trend term lists

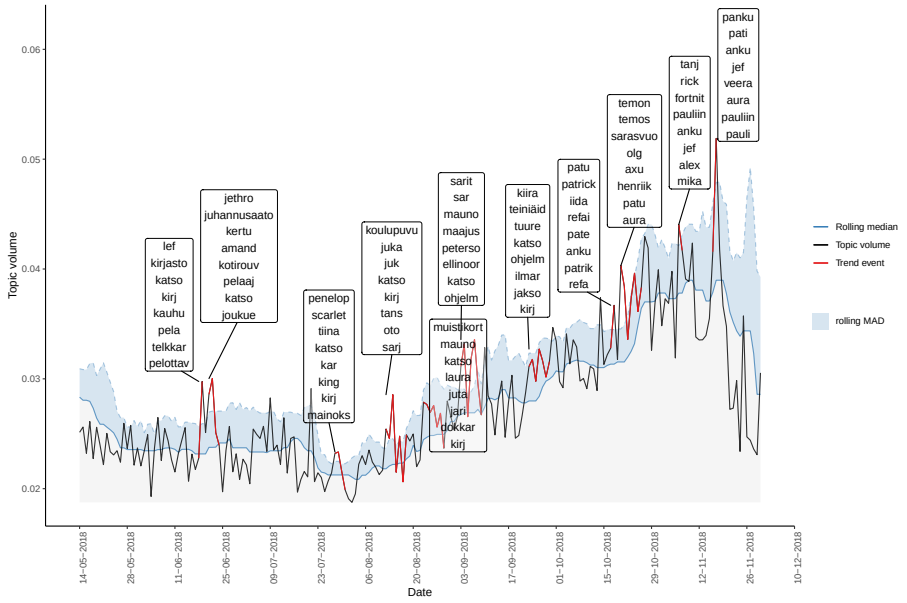


Figure 4.2.3: `vauva.fi` 40 topic LDA model, topic "Media". An interesting feature in this topic is that the trend event terms include a lot of artist and celebrity names, indicating heightened interest or publicity. The upward trend in the topic is also noteworthy as a possible signal for markerting actions.

for each event mostly include the same or similar terms. The terms do not relate to any distinct aspects of the topic but are quite general in the context of the topic. This makes the topic to have low interest in the sense of providing any actionable insight for business or marketing. See Figure 4.2.7 for trend events in the topic "Sex".

Besides the topical trend events identified in the topics, the longer term trend or variation in topic volume can also be studied in the trend event graphs. This information is useful for identification of seasonality in the topics. A great example of this is the topic about "Vacations", which is displayed in Figure 4.2.8. The topic exhibits a clear seasonal increase of discussion volume during July and August. There is also a single large trend event during this period that has trend terms related to heat waves, blue algae and air conditioning. The topic has much lower volume during September (the end of summer vacations) with slight increase towards the end of the study period.

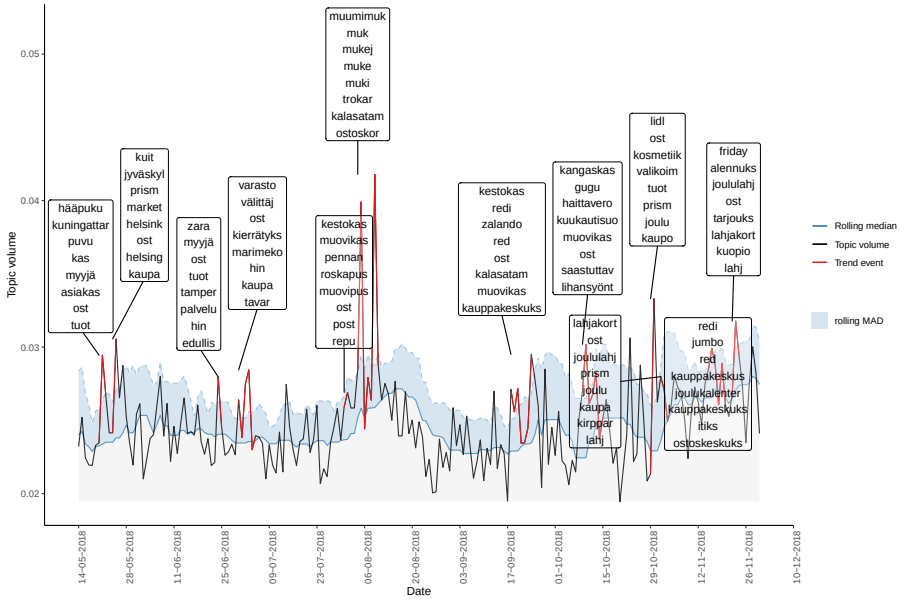


Figure 4.2.4: `vauva.fi` 40 topic LDA model, topic "Shopping". The trend events feature some store brands such as the clothing chain *Zara*, web store *Mr. Gugu*, grocery stores *Prisma* and *Lidl* as well as the shopping malls *Redi*, *Jumbo* and *Itis*.

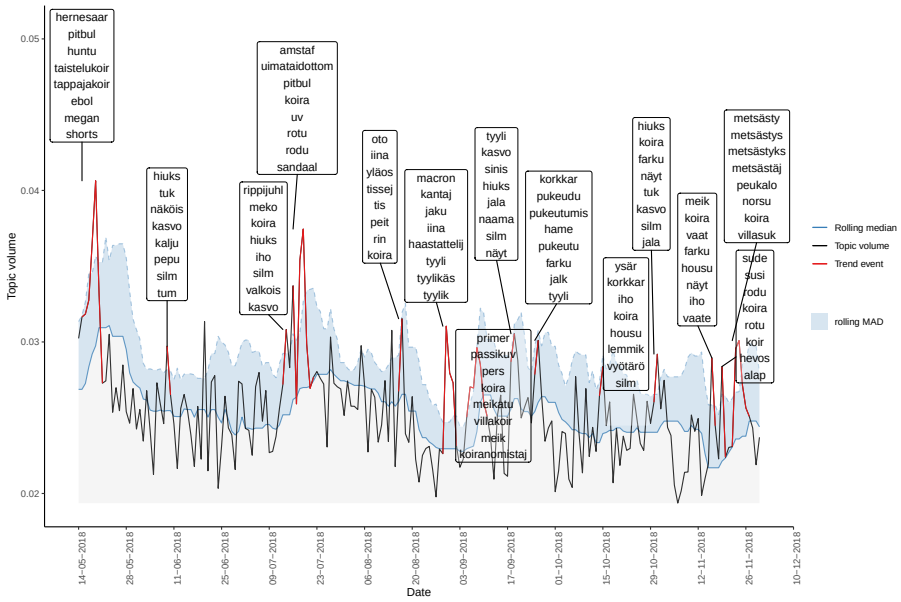


Figure 4.2.5: `vauva.fi` 40 topic LDA model, topic "Looks". Many of the trend events relate to looks, makeup, clothing and such, but there are also four clear trend events that are about animal related hobbies. This indicates that desipte having only one pet related term in the overall top word list (dog), the topic is actually a mix between looks and animal hobbies.

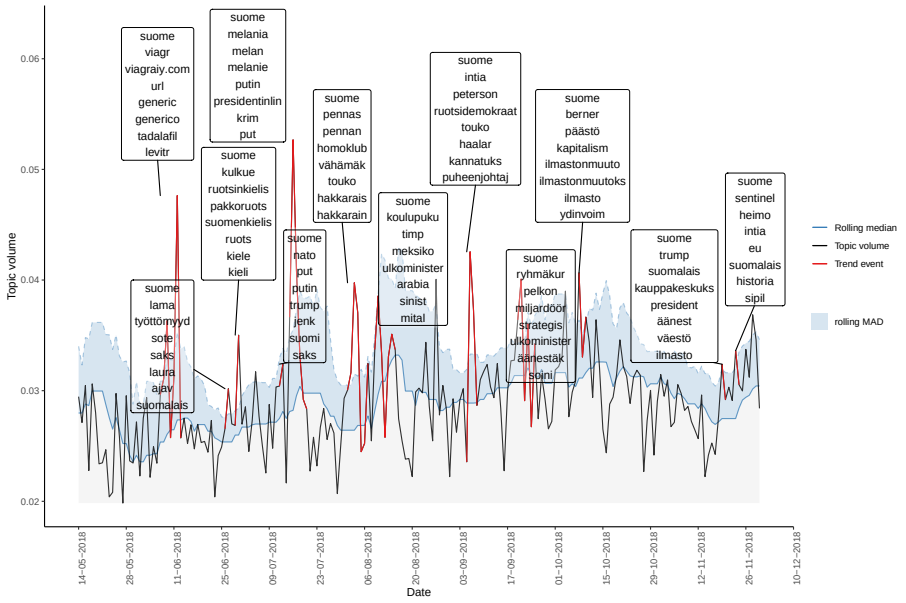


Figure 4.2.6: `vauva.fi` 40 topic LDA model, topic "Politics". The first trend event does not relate well to the topic or the rest of the trend events. Important current affairs are identified, such as the Helsinki summit of presidents Trump and Putin (fourth event, highest topic volume) and how member of the parliament Jaana Pelkonen broke the group discipline in a parliament vote (event in late September).

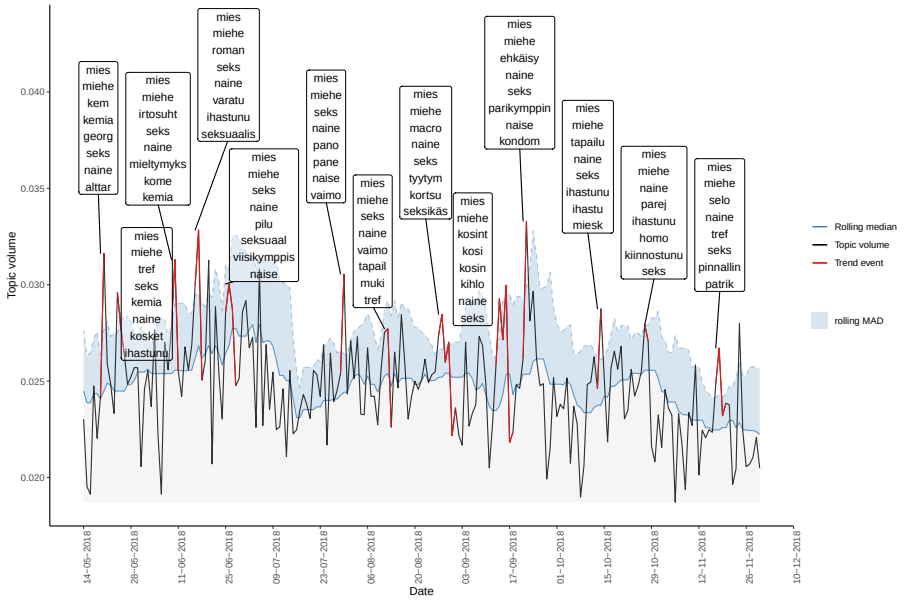


Figure 4.2.7: `vauva.fi` 40 topic LDA model, topic "Sex". The trend events have few terms with relevant information about the distinctive reason behind the trend event. Most terms are general topic related words making this topic of low interest as a source of actionable insights for business or marketing.

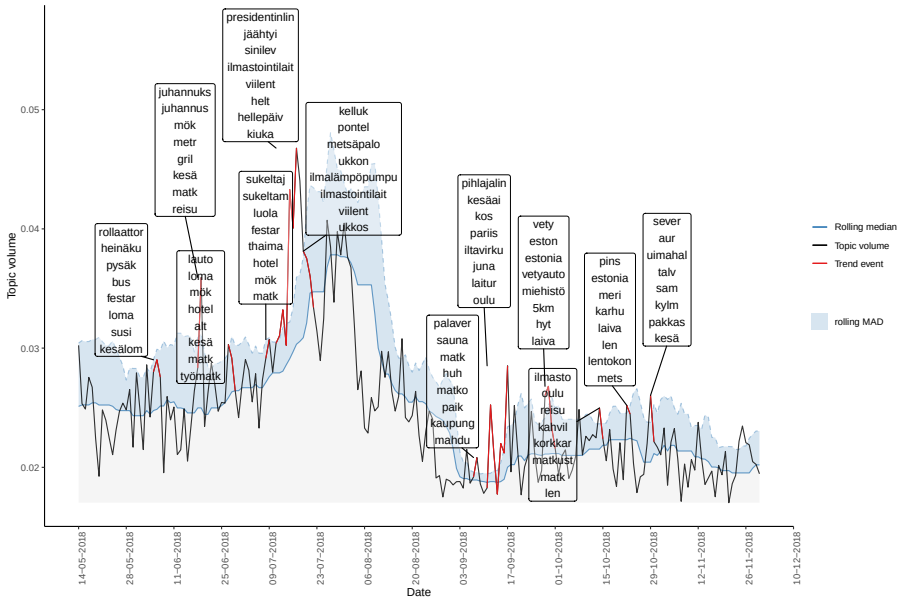


Figure 4.2.8: `vauva.fi` 40 topic LDA model, topic "Vacations". This topic has large seasonal variation in topic volume, which is a useful insight on its own.

Topical trend events in `suomi24.fi` The topical trend identification method shows viability also for the `suomi24.fi` data. However, the topics do not have as much business or marketing use potential as `vauva.fi`. The main two topical areas in `suomi24.fi` are related to politics and religion.

The politics topics trend events contain many real world affairs, which illustrates the functionality of the topical trend identification method. For example, the topics "Domestic Politics" 1 and 2 trend events include the debate about member of the parliament Laura Huhtasaari thesis plagiation suspicions, the widespread strikes in October and the intervention to forcible return of refugees by member of the parliament Aino Pennanen. On closer inspection, it is evident that the topic "Domestic Politics 1" is a more general topic and the topic "Domestic Politics 2" is more about the True Finns political party. The topics and their trend events are displayed in Figures 4.2.9 and 4.2.10. The topic "Climate change" contains trend events related to environmental discussion such as the IPCC climate report and the heat waves of summer 2018 and the subsequent forest wildfires, presented in Figure 4.2.11.

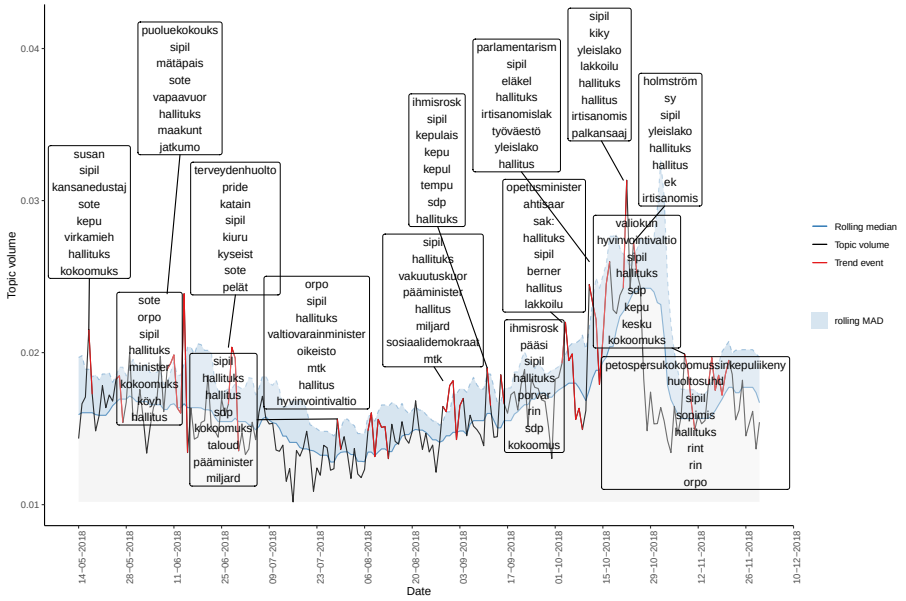


Figure 4.2.9: Suomi24 LDA 60, topic "Domestic Politics 1". Trend events about the general strikes in Finland in Autumn 2018 are identified.

Some `suomi24.fi` topics and trend events may have relevant information for business and marketing. The topic "Cars" contains discussion about car manufacturers and diesel powered cars. The trend events are, however, quite sparse and do not show clear connection to e.g. new car model launches. The information about trending brands might still be relevant. The topic "Personal electronics" was hypothesized to contain very interesting information about customer preferences to electronics manufacturer brands, but the trend events turned out to be more about computer maintenance and debate between fans of different manufacturers (indicated e.g. by the trend words "winhihhuli", loosely translated to a "looney Windows fan" and "arkkivihollinen" or "arch nemesis"). The topical trends for "Cars" and "Personal

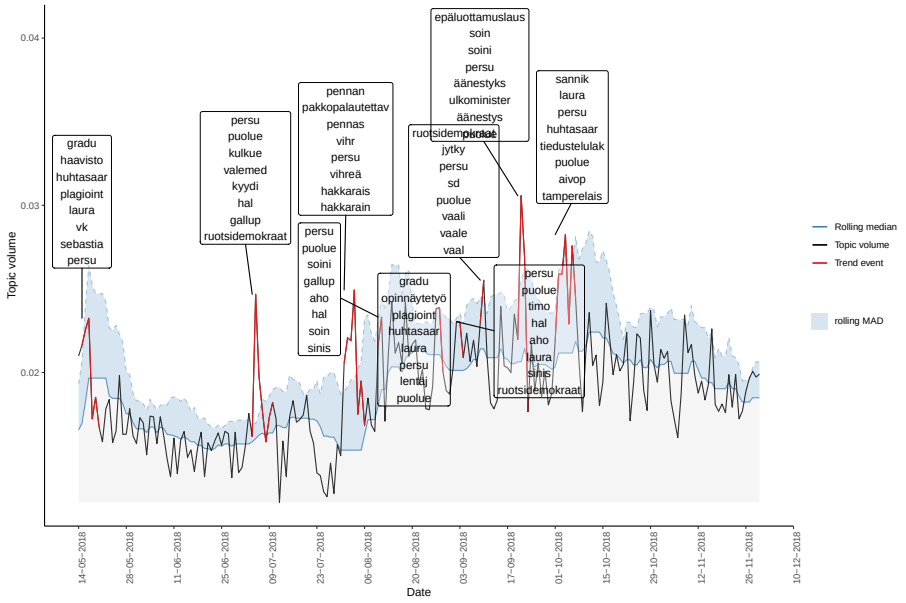


Figure 4.2.10: Suomi24 LDA 60, topic "Domestic Politics 2". On closer inspection, this topic relates mostly to the True Finns political party.

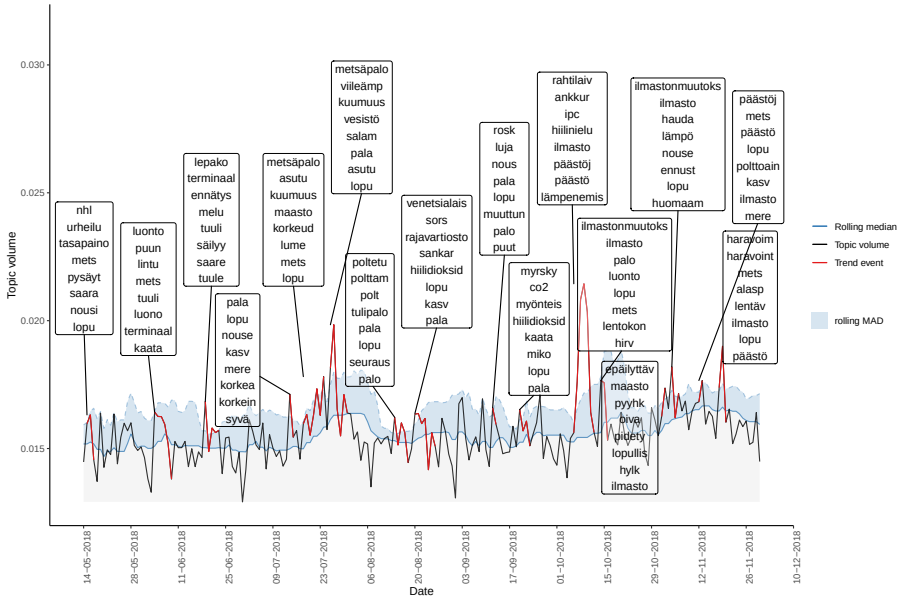


Figure 4.2.11: Suomi24 LDA 60, topic "Climate change". The publication of the IPCC climate report and the forest wildfires of summer 2018 are identified.

electronics" are displayed in Figures 4.2.12 and 4.2.13.

It was pointed out in Subsection 4.1.2, that a notable number of `suomi24.fi` topics related to obscene language use and debate discussion. If such topics would show significant trends in topical volume, they could possibly be used to estimate the overall attitudes of the audience. However, e.g. the topic with a top word list of vulgar trash talk showed little variation in topical volume as seen in Figure 4.2.14.

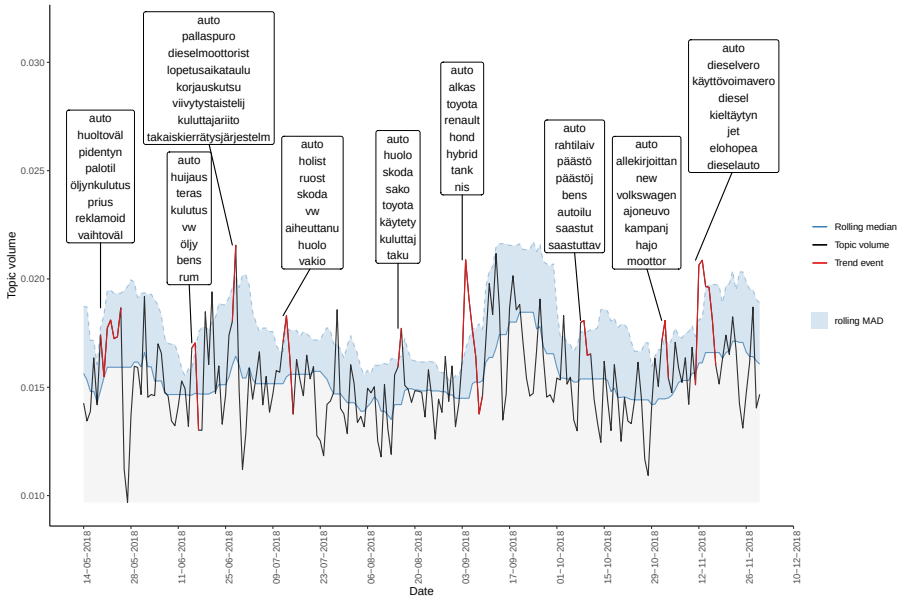


Figure 4.2.12: Suomi24 LDA 60, topic "Cars". The trend events provide insight on manufacturer brands in discussion but do not directly connect to e.g. new car model launches. The discussion about diesel power is important in this topic.

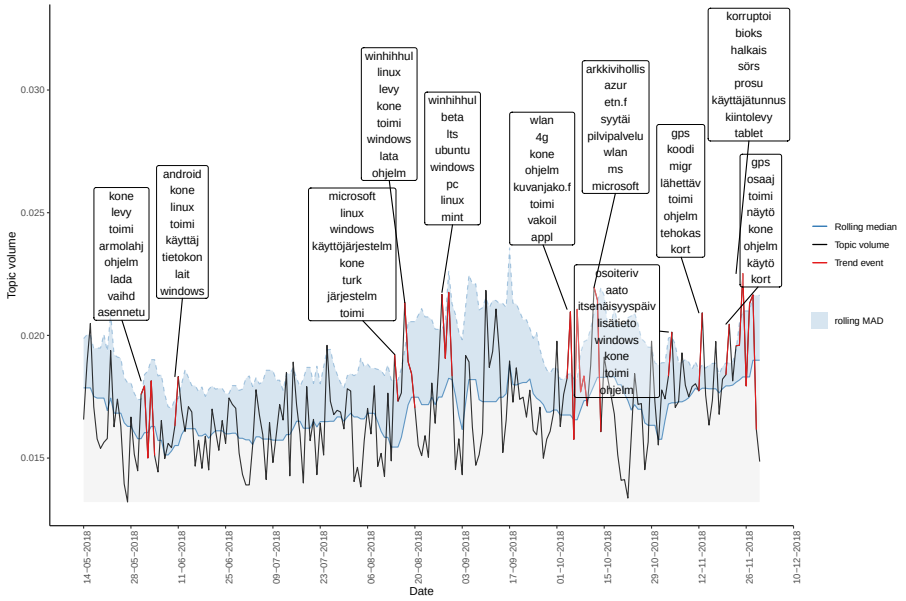


Figure 4.2.13: Suomi24 LDA 60, topic "Personal electronics". The topics was hypothesized to contain information about the customer preferences of electronic devices but the trend events mostly relate to maintenance of computer devices and debate about different brands (e.g. Windows vs. Linux).

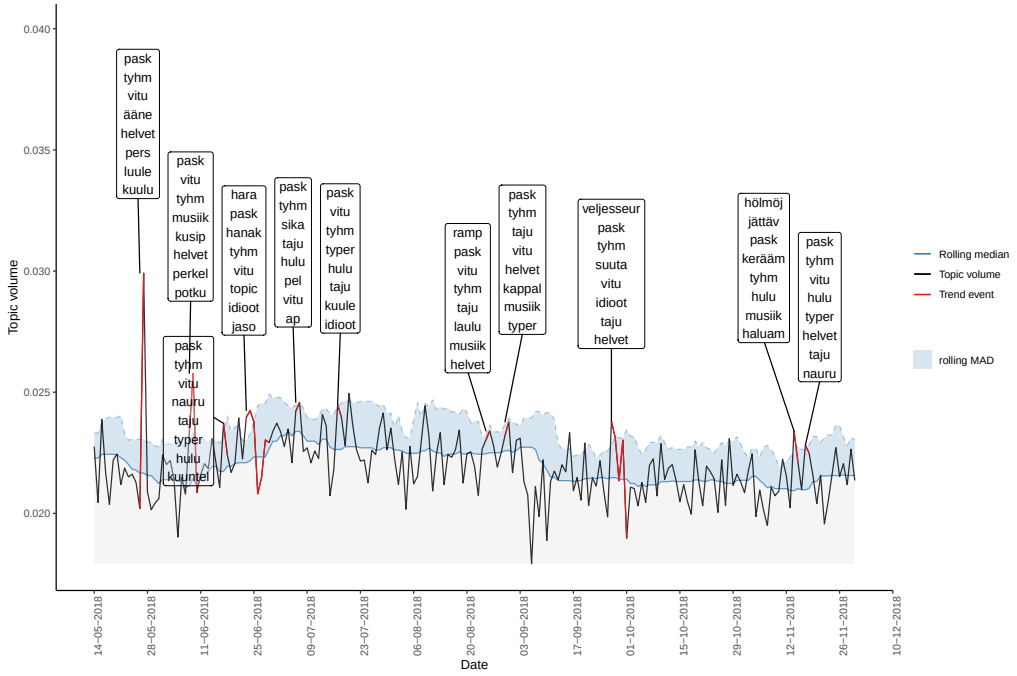


Figure 4.2.14: Suomi24 LDA 60, topic "Trash talk". The volume of vulgar discussion does not exhibit significant temporal variations or meaningful trend events.

4.3 Applications of topical trends in marketing

The data source of this thesis (Finnish Internet discussion forum posts) enables gaining insight on a wide variety of topics concerning the daily news, food and nutrition discussion, media, politics and issues related to our society as well as family, relationships and dating. While these are likely to be useful in some types of business, the data is not optimal for straightforward application in marketing. The results give confidence that the approach presented in this thesis has good performance, given the quality of the source data. Widening the scope of the data sources to more domain specific text data can increase the applicability of the results for marketing. A more specific data source increases the density of the relevant topics and furthermore, relevant trend events. The high quality of the source data is a precondition for high quality results, as is the case in any data analysis. This subsection studies the possible applications, but does by no means collect an exhaustive list of all possible applications. The food industry is used as an example for applying the topic trends in marketing, because it is easily understandable and relevant to everybody, as we all shop groceries on a daily basis.

Possibilities from using more diverse data sources The application method in this thesis is versatile with respect to the data source. Finding trends in discussions is not far from finding trends in any kind of text data, which is hypothesised to contain relevant information. For example, a client in the food industry is interested in the trends in recipes and popularity of diets. In the general discussion forum data, only a fraction of the entire topic space is related to this topic (e.g. 2.7% of the `vauva.fi` topics). Replacing the general discussion forums with a selection of food blogs, competitor recipe sites, social media feeds concerning food and diets, grocery store product wishlists and customer feedback data, makes a far larger portion of the entire topic space to be about food and related subtopics. Another approach could be to first use a primary topic model (such as the one presented in this thesis) to label posts in the general internet discussion forum that are related to an area of interest, and then rerunning the topic model to further split this set of posts into multiple subtopics.

The trend identification method is also applicable not only for text data concerning a large amount of customers but also for more restricted data sets. For example, using customer feedback data of a single business or product as the source makes it possible to track relevant issues by identifying trending topics in real time. The challenges in this type of applications is the size and the quality of the data, making model evaluation a very important step before drawing conclusions.

The application method is not language independent, meaning that it does not deal well with data which contains multiple languages and is likely to split each language into separate topics. However, the method is applicable to any language, given that the entire document space is in the same language. This makes it possible to use e.g. foreign discussion forums, blogs, social media feeds or company web sites as a data source. This is a lucrative possibility for a client to obtain knowledge about trends in a foreign market and drawing conclusions about their possible migration

into the local market. E.g. in the food industry, some of the trends that are current in the United States or in Central Europe might migrate to Finland with some delay. Studying historical data can give insight on the migration speed and scale of the trends.

Using insight from identified topic trends in marketing applications The actual usage of the trend insights has many possible directions. The traditional use of the insights as a source for content creation (written articles, visual creatives) and product planning are the most obvious ones. Staying on the edge on customer needs and preferences creates a competitive advantage for a company, be it launching a new product or marketing existing products reflecting the trends of the customer base.

The presented application method enables a high level of automation of the trend and topic identification through automated data collection (web crawling) and analysis (continuous model and result updates). This provides many application possibilities beyond traditional insight usage. Contemporary marketing automation tools can draw on the topic and trend identification results, effectively removing the slow human component from the marketing pipeline. For example, the identified trends can be used to automatically create search engine marketing keywords for client campaigns. Replacing human generated search keywords with continuously updated keywords from the trend events can increase the reach and relevance of search advertising. For a client in food industry, selecting current food and diet related trends as a basis for the search keywords enables the client to reach audiences interested in the current trends in real time, without manually tweaking the search keywords.

Another automated marketing application is content recommendation. The client might have relevant content (e.g. articles, blog posts, recipes) related to trends, but they have been published a long time ago such that they are not actively presented to the customers on the website. Tagging content with rich and descriptive keywords makes it possible to find content that matches to the current topic trends. Such matching content can then be delivered to the customers increasing the amount of relevant content and interest. Delivery in this case can be e.g. presenting the relevant content as headline articles on a website, increasing matched content rank in the recommended content, increasing media budget for matched social media ads or even including the matched content as the landing pages in search or display adverts. With the use of contemporary marketing automation tools, the matched content can even be used as the content for an advert (visual and copy text creatives). Content matching can also be done for visual creatives. Tagging a database of visual resources and matching with topic trends can be used in automatic creative generation. Instead of tagged, pre-made visual creatives, graphical content generated by artificial neural networks could also be used. *General Adversarial Networks* have shown promise in visual content generation based on keywords. The benefit of using automatically identified trends and topics in all of the proposed applications in this paragraph, is the removal of the slow human component in finding and curating trends.

In addition to the topical trends, also the basic topic modelling results can be

used. For example, moderating or structuring comment sections on a website or a discussion forum may benefit from the use to topic models as a source of information. If a human user first identifies topics of low interest or topics, which connect to abusive content, posts that relate to these topics may be deleted or discouraged.

While it was previously stated that the density of business relevant topics in the general Internet discussion forums is a shortcoming, it is also possible to use the more general discussion themes about politics and society in marketing. The information of popular concerns about current affairs can be used in risk management of marketing as well as in the consideration the potential channels where to raise higher customer interest. For example, the method in this thesis identified the strike of the Finnish public and welfare worker's union (JHL) as a trend event in the discussion forums. Gaining this type of information in real time can help companies to avoid marketing with a message that crosses the public opinion about an issue like this. Using again the example of the food industry, a company could plan a campaign that addresses the public discontent on the strike by providing free meal samples etc. to the audiences that have been affected by the strike. The benefit of the automated trend identification in this case does not come from automation but the reaction speed advantage of real time identification speed. Another way to utilize the information of trends in current affairs is to use media channels that deliver content on the trending affair. E.g. the final episodes of the Finnish TV reality show "Love Island Suomi" was identified as a trend in `vauva.fi` topic "Media", so delivering adverts on media, which report on the TV show, could be a beneficial way to reach a larger audience.

Summary of applications in marketing Table 4.3.1 summarizes the different application possibilities outlined in this subsection. The list nor this subsection are intended to be exhaustive of all possible applications. All applications may benefit from enriching the results by the use of more domain specific and varied text data sources e.g. blogs, articles, social media feeds or company websites.

Table 4.3.1: A summary of possible marketing applications of the topic trend identification method presented in this thesis. The list is not intended to be exhaustive of all possible applications.

	Application
-	Manual use of trend and topic insights, e.g. for researchers or content creators
-	Trend insight for product design and development
-	Trend identification from foreign market text data and estimation of trend migration to local market
-	Automatic SEM keyword list generation based on identified trends
-	Automatic SEO (Search engine optimization): website content and meta information updating based on identified trends
-	Automatic (pre-existing) content matching based on identified trends and their suggestion or delivery to users
-	Automatic content creation based on tagged visual resources and matching to identified trends
-	Automatic moderation based on topics related to abusive or low interest themes
-	Focused delivery media channels that reach audiences interested in identified trends
-	Risk management of ad or media delivery, based on identified trends with negative connotations

5 Summary and future prospects

5.1 Future prospects

The topical trend identification application presented in this thesis provided promising results. The method was able to identify realistic trend events from Finnish Internet discussion forum data. There are multiple ways to further improve the method. For the future prospects of applying the topic trend identification in marketing see Subsection 4.3.

Data preprocessing Data preprocessing has a large effect on the quality of the results and also comprises a substantial part of the method's run time. In the application of this thesis, data preprocessing cut some corners with respect to state of the art, which provides natural directions for further development.

The lemmatization, or alternatively stemming of the terms, is an important phase in BOW model data preprocessing. The Finnish language has especially complicated morphology, necessitating the use of either lemmatization or stemming in order to have a reasonably sized vocabulary. Including all possible morphologies of basic words would increase the size of the vocabulary by an order of magnitude. In this thesis, stemming was used instead of lemmatization. Stemming is easy to implement with readily available methods such as the *SnowballC* -stemmer used in this thesis. However, stemming is not able to detect the true stem of the words with 100% accuracy, resulting in multiple instances of the same basic word. This is seen e.g. in the top word list of the largest `vauva.fi` topic about politics where the word "Finnish" occurs multiple times as "suome", "suomalais", "suomi" and "suomalain" which all have the same stem.

Replacing stemming with lemmatization would have multiple benefits. The vocabulary would have only true lemmas and thus a reduced amount of duplicate words, making the analysis computationally easier. The use of lemmatization would also remove the trouble from the user to make a distinction between different stemmed versions of the same lemma. In addition, the stemmed words are sometimes difficult to interpret and make the results harder to read. Another advantage of lemmatization would be to know the part of speech (POS) of each term and subsequently to prune the vocabulary based on this. For example verbs are not necessarily useful for the topic model analysis, because their meaning is highly related to the context, whereas nouns have far greater contribution on the meaning of the text. This would allow for better functioning vocabularies to be used in the topic modelling.

Lemmatization can be done by using tools included in some NLP software libraries such as the the Natural Language Toolkit (nltk) for Python. The selection of the languages and morphology sources is limited in these libraries and e.g. nltk does not have a Finnish lemmatizer. There are some Finnish lemmatizer projects which could be used in this applications: The TurkuNLP parser pipeline (Kanerva et al., 2018) and Helsinki University supported Open Morphology for Finnish (OMORFI) project maintainted in GitHub under *omorfi*. However, implementing either of these lemmatizers would require significant software integration work, especially

considering the general usage of R pipeline and libraries in this thesis and the lack of R libraries in the aforementioned projects. The use of lemmatization would require overcoming limitations set by the large variance of language and words in the Internet discussion forums compared to the dictionary language. Slang words, abbreviations, product names, words with foreign origin etc. are important for the analysis, but are complicated to handle with the lemmatizers.

Another way to further enrich the scope of the analysis would be to consider using *bigrams* or *n-grams* instead of single words in the vocabulary. An n-gram is a sequence of n contiguous terms in a text and bigram is a set of two terms appearing after one another in a text. This would enable a larger degree of context to be taken into account, but on the other hand would increase the number of terms in the model significantly requiring a great boost in computational resources compared to the setup used in this thesis. Another problem can arise from the sparsity of n-grams compared to single words, which would make the models possibly noisier.

Topic models The tested topic models, the LDA and the DTM, were studied to some extent in this thesis. However, much more research especially about the DTM, could be done in the same context as this thesis. A longer period of time and experimentation with different lengths of time slicing in the DTM could give more insight to the behaviour of topic mixtures over time in this kind of text data. A more structured test framework, which would allow direct comparison between LDA and DTM models, would be beneficial for quantitative analysis. This applies also to human evaluation, which was not conducted on the DTM topic model results. The possibility of discovering long term trends directly from the topic mixtures should be studied. This approach would also give a point of comparison for the trend event terms used in this thesis.

The topic models LDA and DTM were used in this thesis but many others were presented in Subsection 2.2. It would be interesting to see and compare the performance of the trend identification method with some other topic models. The additional utility in using other topic models could arise in the possibility of taking topical correlations or hierarchical structure of topics into account. Models such as the CTM, PAM or HDP topic model could be useful. The benefit of using LDA derived models is that the result data format would remain the same, allowing identical method for topic evaluation and trend identification to be used. The use of a PAM model introduces the possibility to model both super-topics and their respective sub-topics, even in more than two levels of hierarchy. This could be helpful in the case of a data source that covers a wide variety of different subjects, such as the Internet discussion forum data, as the topics would be organized in levels of hierarchy. Identifying interesting super-topics would allow the results to be limited to the sub-topics most connected to the interesting super-topic. In addition to modelling a nested correlation structure of topics, HDP would allow the model topic number to be determined by the model, removing the need to manually tune the topic number hyperparameter. More complicated artificial neural network topic models could also be used, but the evaluation and trend identification methods would need to be adjusted according to the model output format. The quality of different

topic models could be evaluated by comparing multiple different topic models fitted on the same data and studying, whether all models exhibit similar results both with respect to the topics as well as the subsequent trend events.

Recently, research about topic models based on artificial neural networks has been conducted. A short review about these methods is in the end of Subsection 2.2. While application of neural network topic models might be difficult before further research, it can be speculated that this kind of approach can have benefits for the application presented in this thesis. If a RNN type topic model could be applied on a *character level*, the complicated tokenization and lemmatization steps in the data preprocessing could be skipped. The model would then learn not only the topical relationships of the text but also the character and word level features of the language. This would potentially require a very large training data but could then forego the challenges of slang words and language dependency related to lemmatization.

A complicated problem in applying the topical trend identification method to real world cases lies in updating the model over time. Inferencing the model with new data is straightforward and enables the tracking of topic volumes and identifying trend events as with historical data. However, updating the entire model, while still retaining the topic structure similar to the previous models, is difficult with the software implementation and libraries used in this thesis. The need to update the model arises, if a lot of new important terms are introduced in the data, but not found in the model or if the topic structure significantly changes over time. A natural solution to this problem of continuously updating the model fit would be to use the current model as a priori information for fitting the model with new data. This is not currently possible with e.g. the *text2vec* R-package, which implements the *WarpLDA* inference method.

Trend identification method The trend identification method developed in this thesis has many possibilities for further development. The method was somewhat heuristic and it was not directly based on pre-existing methods for outlier detection. One challenge with the approach was the manual tuning of the identification threshold coefficient. A possible way to improve suitable threshold coefficient tuning is to annotate trend events manually in the topic volume time series data and to learn the coefficient from the annotated data. Problems arise from the static coefficient (time invariant and shared across all topics) used in the application. Allowing the threshold to be tuned by topic could further improve the trend identification results. However, manual annotation requires a lot of work from several annotators and careful planning to avoid bias.

Annotating trend events manually enables the use of completely other trend event identification approaches, such as classification with recurrent neural networks (RNN) or convolutional neural networks (CNN). The usage of artificial neural networks requires large amounts of data and representation from all different settings the method is planned to be applied on.

Unsupervised alternatives for trend identification include the Kalman filter that could be used to extract noise from the topic volume time series and then to identify the outlying trend events. However, the use of Kalman filters would require

assumptions about the distribution of the topic volume time series residual. In this thesis, there were no assumptions on the distribution, as the distribution free parameters median and median absolute deviation were used.

Interpretation of the trend events relied on finding the important terms of the topic from the period of the identified trend event. This was achieved by calculating the trend lift and subsequent trend relevance as defined in Subsection 3.5.4. The relevance metric could be replaced by other metrics such as the *FREX* metric or pure lift (see Subsection 2.3.1). The weighting between trend lift and relevance in calculating the trend relevance has a significant effect on the resulting list of trend event terms. Systematically evaluating the coherence and connection to real world events of the trend event terms with different weighting schemes and then learning the optimal weights could increase the quality of the trend identification method.

Other future prospects The results of the topic models and the trend events are presented with various static (non-interactive) graphs in this thesis. The main findings are highlighted and analysed, but an interested user has no access to seeing e.g. the trend terms of arbitrary dates. Because the trend identification method produces such large results (averaging over ten trend events per topic with 40 topics in single data set), the most important future prospect is to develop interactive methods for browsing the results and drawing insights. Allowing the user to focus on the topics that interest them the most is imperative due to the large number of trend events in the entire data. Automatic summarization and prioritising of the results would be another lucrative direction for further development. Simple heuristic rules could be enough to make it easier to browse the results, such as giving a list of the most recent trend events in the order of decreasing trend event length or summed proportional residual over the trend identification threshold. Many existing topic model visualization methods could be used to make result interpretation easier (cf. Subsection 2.3.2).

5.2 Summary

The goal of this thesis was to study the use of topic models as a tool to identify trends from Finnish Internet discussion forum text data and to explore the kind of applications the identified trends can have in marketing. Topic models and their evaluation methods were extensively reviewed as a basis for the topical trend identification application. Discussion data was collected from two popular Finnish Internet discussion forums, `www.suomi24.fi` and `www.vauva.fi`. Latent Dirichlet Allocation topic models and Dynamic Topic Models were fitted and topical trends were identified using a method that builds on outlier detection techniques. The results are promising as many real world events could be identified from the discussion data in an unsupervised manner. Possible applications of the topical trends in marketing were studied and future prospects considering the improvement of topic model methods, the trend identification method and possible alternative data sources were presented.

Review of topic models and text analytics in marketing Topic models aim to decompose the unstructured text data into the underlying topics the text represents. Multiple different models first applied as methods for information retrieval have been researched during the 2000s. Latent Dirichlet Allocation topic model, developed by Blei et al. (2003), is the most widely applied topic model and it was also used in the application of this thesis. Building on LDA, other topic models that are able to capture more complex features of the text data have been researched since the introduction of LDA in 2003 (Blei and Lafferty, 2005; Li and McCallum, 2006; Teh et al., 2005). Topic models that allow external variables to be taken into account have also been developed (Blei and Lafferty, 2006; Roberts et al., 2014). A review of LDA-derived topic models by Blei (2012) provides a useful summary of the different methods. Inference of probabilistic topic models is a substantial computational task and different efficient estimation methods have been researched, such as the WarpLDA implementation used for LDA fitting in this thesis (Chen et al., 2016).

Interpretation of the latent or underlying topics of text poses a challenge as the word distributions of the topics are not necessarily meaningful for a human user. The quality of the topic model can be evaluated with different approaches (Wallach et al., 2009; Chang et al., 2009; Newman et al., 2010; AlSumait et al., 2009). Multiple metrics to estimate the importance of different terms in a topic have been researched (Taddy, 2012; Bischof and Airoldi, 2012; Chuang et al., 2012a; Sievert and Shirley, 2014). The *relevance* metric proposed by Sievert and Shirley (2014) was used to to evaluate term importance in the application in this thesis. In addition to plain lists of important terms in a topic, labelling topics for easier interpretation has been researched (Lau et al., 2011; Zhao et al., 2011). The visualization of the topic modelling results can be achieved by various methods (Chaney and Blei, 2012; Chuang et al., 2012a; Sievert and Shirley, 2014).

Text analytics have seen an increasing use in marketing, both in scientific research and in businesses. Different uses include media tracking, sentiment analysis, text categorization and semantics, chatbots and automatic discussion moderation. See

Subsection 2.4 for a review.

Identification of topical trends in Finnish Internet discussion forums Finnish Internet discussion forums `www.suomi24.fi` and `www.vauva.fi` were used as the source data for the application. Data was collected by using a web crawler that scraped data from the forums from May 2018 to November 2018, totalling 4 189 313 individual forum posts. LDA models were fitted using the WarpLDA inference method implemented in the `R` package `text2vec`. Evaluation of a suitable topic number was conducted using both model perplexity and human evaluation. The `suomi24.fi` data was decomposed into 60 topics and `vauva.fi` data into 40 topics, which provided the best human interpretable results. A total of 37 topics in `suomi24.fi` and 30 topics in `vauva.fi` were interpreted to have a coherent semantic meaning. For example, topics about relationships, child care, politics and food were discovered. The topics in `vauva.fi` had more relevance for marketing and business use, as a large portion `suomi24.fi` topics comprised debate about religion and politics. A Dynamic Topic Model of 30 topics with one week time slices was also fitted on a sample of 500 000 posts from both forums. The DTM was observed to exhibit little dynamic variations in the topic term mixtures, which justified the use of the more simple and efficient LDA topic model for further results.

Topical trend identification was based on examining the daily proportional topic volumes in the discussion data. A method that finds short time intervals of extraordinarily high proportional topic volume was developed and used to identify trend events. Terms that describe each separate trend event were found by comparing relative term importance during the trend event and the entire data as well as taking into account the term relevance for the topic in question. The trend events and the trend terms were compared against real world events to validate the quality of the trend identification method. Both `suomi24.fi` and `vauva.fi` topical trend events were observed to have a connection to real world, but as in the general topic model results, `vauva.fi` trend events provided more relevant information for marketing and business use. Examples of identified events included the Finnish high school graduation and related celebration dishes, the final episodes of the TV reality show "Love Island Suomi" and its participants, the launch of a rare *Moomin mug* by Iittala, the opening of the *Redi* shopping mall and a general increase in discussion about vacations during the summer. These events are examples of relevant signals for marketing. In addition, events about e.g. the general strikes in Finland, the IPCC climate report and the summit of presidents Putin and Trump were identified but the relevance of these events for marketing is not as significant.

Applications in marketing The identified topical trends can have different uses in marketing. Since the trend results can be obtained in an automated fashion by applying the pipeline of data collection, preprocessing, topic modelling and trend identification, the applications also reflect the use of automation and removal of manual work. Topical trends can be used as an input for e.g. automatic search engine marketing keyword list generation, website search engine optimization, matching pre-existing content to trends for content recommendation on websites and adverts

as well as automatic generation of visual creatives by matching trend terms to visual resource tags. Besides the uses in automating marketing actions, topical trends can be utilized for more manual applications. Using the presented topical trend identification for foreign market data sources can provide insight on future trends and their migration speed to a domestic market. The trend insight can be used as a plain source of information for market researchers, content creators or product owners. Applications are studied in Subsection 4.3.

In conclusion, the studied Finnish Internet discussion forums `suomi24.fi` and `vauva.fi` could be used to identify current events and trends in Finland. Software implementations for topic modelling are readily available and data can be obtained from public Internet sources. The results are promising and applying the method in more domain specific data sources could provide even more relevant trend insights. Further development of the methods presented in this thesis is seen worthwhile as the possible marketing applications have genuine use in business.

References

Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Alexander, E. and Gleicher, M. (2016). Task-driven comparison of topic models. *IEEE transactions on visualization and computer graphics*, 22(1):320–329.

AlSumait, L., Barbará, D., Gentle, J., and Domeniconi, C. (2009). Topic significance ranking of lda generative models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 67–82. Springer.

Andrzejewski, D., Zhu, X., and Craven, M. (2009). Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th annual international conference on machine learning*, pages 25–32. ACM.

Atherton, P. and Borko, H. (1965). A test of factor-analytically derived automated classification methods. *AIP rept AIP-DRP*, pages 65–1.

Baker, F. B. (1962). Information retrieval based upon latent class analysis. *Journal of the ACM (JACM)*, 9(4):512–521.

Bischof, J. and Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 201–208.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M. and Lafferty, J. D. (2005). Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, pages 147–154. MIT Press.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

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

Borko, H. (1963). Automatic document classification. *Journal of the ACM*, 10:151–162.

Boyd-Graber, J., Blei, D., and Zhu, X. (2007). A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Bray, T. (2017). The javascript object notation (json) data interchange format.

Bush, V. (1945). As we may think. *The atlantic monthly*, 176(1):101–108.

Chaney, A. J.-B. and Blei, D. M. (2012). Visualizing topic models. In *ICWSM*.

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

Chen, J., Li, K., Zhu, J., and Chen, W. (2016). Warplda: a cache efficient o (1) algorithm for latent dirichlet allocation. *Proceedings of the VLDB Endowment*, 9(10):744–755.

Chen, W.-Y., Chu, J.-C., Luan, J., Bai, H., Wang, Y., and Chang, E. Y. (2009). Collaborative filtering for orkut communities: discovery of user latent behavior. In *Proceedings of the 18th international conference on World wide web*, pages 681–690. ACM.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.

Chomsky, N. (1957). Syntactic structures.

Chuang, J., Gupta, S., Manning, C., and Heer, J. (2013). Topic model diagnostics: Assessing domain relevance via topical alignment. In *International Conference on Machine Learning*, pages 612–620.

Chuang, J., Manning, C. D., and Heer, J. (2012a). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the international working conference on advanced visual interfaces*, pages 74–77. ACM.

Chuang, J., Ramage, D., Manning, C., and Heer, J. (2012b). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452. ACM.

Cui, W., Liu, S., Wu, Z., and Wei, H. (2014). How hierarchical topics evolve in large text corpora. *IEEE transactions on visualization and computer graphics*, 20(12):2281–2290.

Davis, M. and Edberg, P. (2018). *Unicode Emoji 11.0*. Unicode Consortium. Accessed: 2018-07-18.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Dieng, A. B., Wang, C., Gao, J., and Paisley, J. (2016). Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702*.

Dostert, L. E. (1955). The georgetown-ibm experiment. *1955). Machine translation of languages. John Wiley & Sons, New York*, pages 124–135.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285. Acm.

Eeuwen, M. v. (2017). Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers. Master's thesis, University of Twente.

Ester, M., Kriegel, H.-P., and Schubert, M. (2004). Accurate and efficient crawling for relevant websites. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 396–407. VLDB Endowment.

Gardner, M. J., Lutes, J., Lund, J., Hansen, J., Walker, D., Ringger, E., and Seppi, K. (2010). The topic browser: An interactive tool for browsing topic models. In *NIPS Workshop on Challenges of Data Visualization*, volume 2. Whistler Canada.

Goldstone, A. (2016). Data for research browser. http://agoldst.github.io/dfr-browser/. Accessed: 2018-11-01.

Google (2010). Google Cloud Storage: Online data storage. https://cloud.google.com/storage/. Accessed: 2018-07-18.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

Griffiths, T. L. and Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 24.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 363–371. Association for Computational Linguistics.

Hayes, P. J. and Weinstein, S. P. (1990). Construe/tis: A system for content-based indexing of a database of news stories. In *IAAI*, volume 90, pages 49–64.

Hinton, G. E. and Salakhutdinov, R. R. (2009). Replicated softmax: an undirected topic model. In *Advances in neural information processing systems*, pages 1607–1614.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Hoque, E. and Carenini, G. (2015). Convisit: Interactive topic modeling for exploring asynchronous online conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 169–180. ACM.

Hutchins, J. (1997). Fifty years of the computer and translation. *Machine Translation Review*, 6(1997):22–24.

Jardine, N. and van Rijsbergen, C. J. (1971). The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, 7(5):217–240.

Jockers, M. L. and Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, 41(6):750–769.

Johnson, M. (2009). How the statistical revolution changes (computational) linguistics. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 3–11. Association for Computational Linguistics.

Jones, K. S. (1994). Natural language processing: a historical review. In *Current issues in computational linguistics: in honour of Don Walker*, pages 3–16. Springer.

Kanerva, J., Ginter, F., Miekka, N., Leino, A., and Salakoski, T. (2018). Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 133–142.

Kralj Novak, P., Smailović, J., Sluban, B., and Mozetič, I. (2015). Sentiment of emojis. *PLOS one*, 10(12).

Krestel, R., Fankhauser, P., and Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 61–68. ACM.

Lacoste-Julien, S., Sha, F., and Jordan, M. I. (2009). Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems*, pages 897–904.

Lau, J. H., Grieser, K., Newman, D., and Baldwin, T. (2011). Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1536–1545. Association for Computational Linguistics.

Lee, H., Kihm, J., Choo, J., Stasko, J., and Park, H. (2012). ivisclustering: An interactive visual document clustering via topic modeling. In *Computer graphics forum*, volume 31, pages 1155–1164. Wiley Online Library.

Li, W. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. ACM.

Lienou, M., Maitre, H., and Datcu, M. (2010). Semantic annotation of satellite images using latent dirichlet allocation. *IEEE Geoscience and Remote Sensing Letters*, 7(1):28–32.

Liu, B., Liu, L., Tsykin, A., Goodall, G. J., Green, J. E., Zhu, M., Kim, C. H., and Li, J. (2010). Identifying functional mirna–mrna regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics*, 26(24):3105–3111.

Liu, S., Wang, X., Chen, J., Zhu, J., and Guo, B. (2014). Topicpanorama: A full picture of relevant topics. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 183–192. IEEE.

Liu, S., Zhou, M. X., Pan, S., Song, Y., Qian, W., Cai, W., and Lian, X. (2012). Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2):25.

Luhn, H. P. (1958). *Auto-encoding of documents for information retrieval systems*. IBM Research Center.

Mcauliffe, J. D. and Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems*, pages 121–128.

McCarthy, E. J. (1964). *Basic marketing: a managerial approach*. RD Irwin.

Mei, Q., Shen, X., and Zhai, C. (2007). Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499. ACM.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics.

Miner, G., Elder IV, J., and Hill, T. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.

Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*.

Newman, D., Asuncion, A., Smyth, P., and Welling, M. (2009). Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10(Aug):1801–1828.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

Ossorio, P. G. (1966). Classification space: A multivariate procedure for automatic document indexing and retrieval. *Multivariate Behavioral Research*, 1(4):479–524.

Ourila, J. (2018). Fiam - finnish internet audience measurement. http://fiam.fi/tulokset/. Accessed: 2018-07-30.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., and Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis.

Pinoli, P., Chicco, D., and Masseroli, M. (2014). Latent dirichlet allocation based on gibbs sampling for gene function prediction. In *Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on*, pages 1–8. IEEE.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Purver, M. (2011). Topic segmentation. *Spoken language understanding: systems for extracting semantic information from speech*, pages 291–317.

Rackley, J. (2015). *Marketing Analytics Roadmap*. Springer.

Ramage, D., Dumais, S. T., and Liebling, D. J. (2010). Characterizing microblogs with topic models. *ICWSM*, 10(1):16.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics.

Reifler, E. (1954). The first conference on mechanical translation. *Mechanical Translation*, 1(2):23–32.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Scrapinghub (2008). Scrapy: Open source web crawling framework. https://doc.scrapy.org. Accessed: 2018-07-16.

Selivanov, D. (2016). text2vec: Modern text mining framework for r. *Computer software manual](R package version 0.4. 0). Retrieved from https://CRAN. R-project. org/package= text2vec*.

Sievert, C. and Shirley, K. (2014). Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.

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

Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230.

Taddy, M. (2012). On estimation and selection for topic models. In *Artificial Intelligence and Statistics*, pages 1184–1193.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392.

Teh, Y. W., Newman, D., and Welling, M. (2007). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1353–1360.

Turing, A. (1950). Computing machinery and intelligence-. *Mind*, 59(236):433.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning*, pages 1105–1112. ACM.

Wang, C., Blei, D., and Heckerman, D. (2008). Continuous time dynamic topic models. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 579–586. AUAI Press.

Ward, I. (2017). JSON Lines text file format. http://jsonlines.org/. Accessed: 2018-07-18.

Weaver, W. (1955). Translation. *Machine translation of languages*, 14:15–23.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

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

Yao, L., Mimno, D., and McCallum, A. (2009). Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946. ACM.

Yu, H.-F., Hsieh, C.-J., Yun, H., Vishwanathan, S., and Dhillon, I. S. (2015). A scalable asynchronous distributed algorithm for topic modeling. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1340–1350. International World Wide Web Conferences Steering Committee.

Yuan, J., Gao, F., Ho, Q., Dai, W., Wei, J., Zheng, X., Xing, E. P., Liu, T.-Y., and Ma, W.-Y. (2015). Lightlda: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1351–1361. International World Wide Web Conferences Steering Committee.

Zhao, W. X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E.-P., and Li, X. (2011). Topical keyphrase extraction from twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 379–388. Association for Computational Linguistics.

A Trend events in `suomi24.fi`

This appendix lists the figures of topical trend volumes and identified trend events for `suomi24.fi` LDA model with 60 topics. Only topics that were evaluated to have a coherent semantic meaning based on the top 10 terms by relevance, are listed. Refer to Figure 4.1.6 for the relative sizes and top words for each topic.

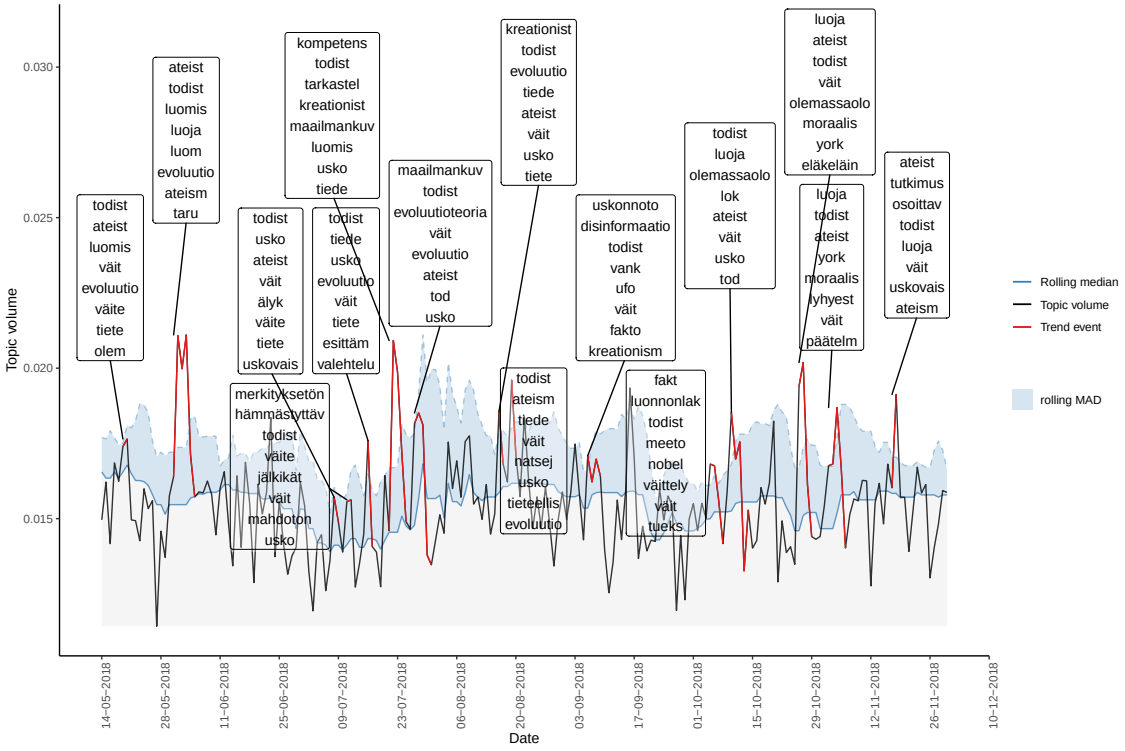


Figure A1: Suomi24 LDA 60, topic "Atheism".

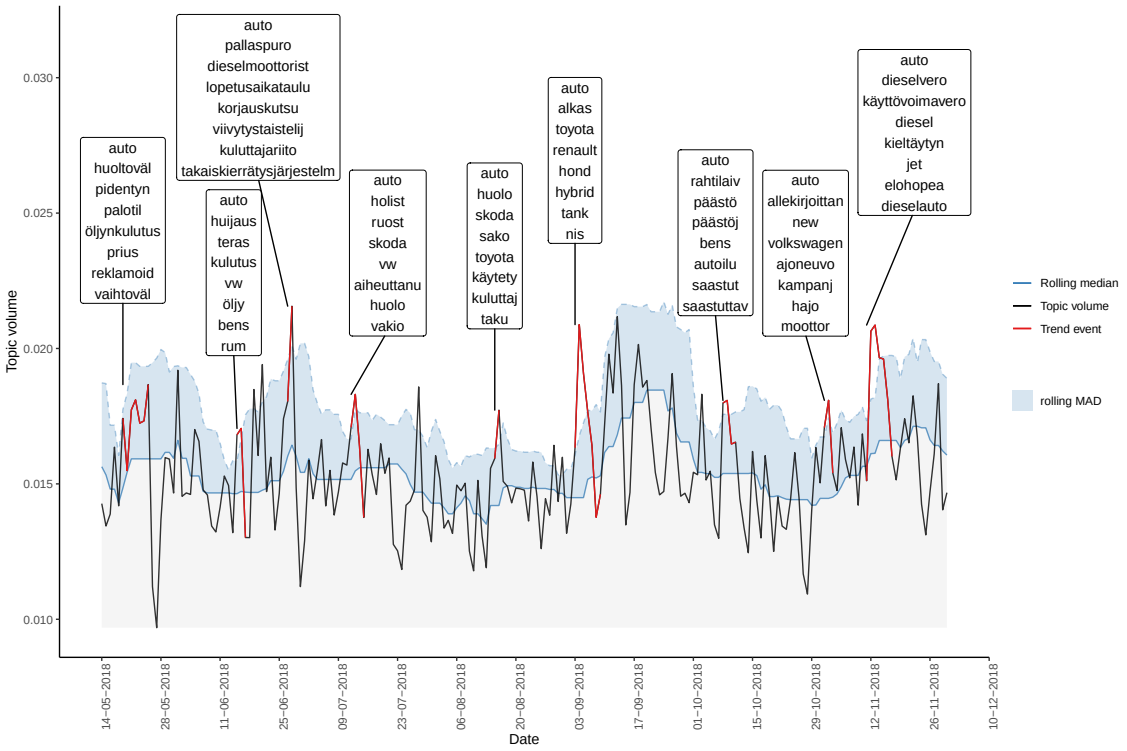


Figure A2: Suomi24 LDA 60, topic "Cars".

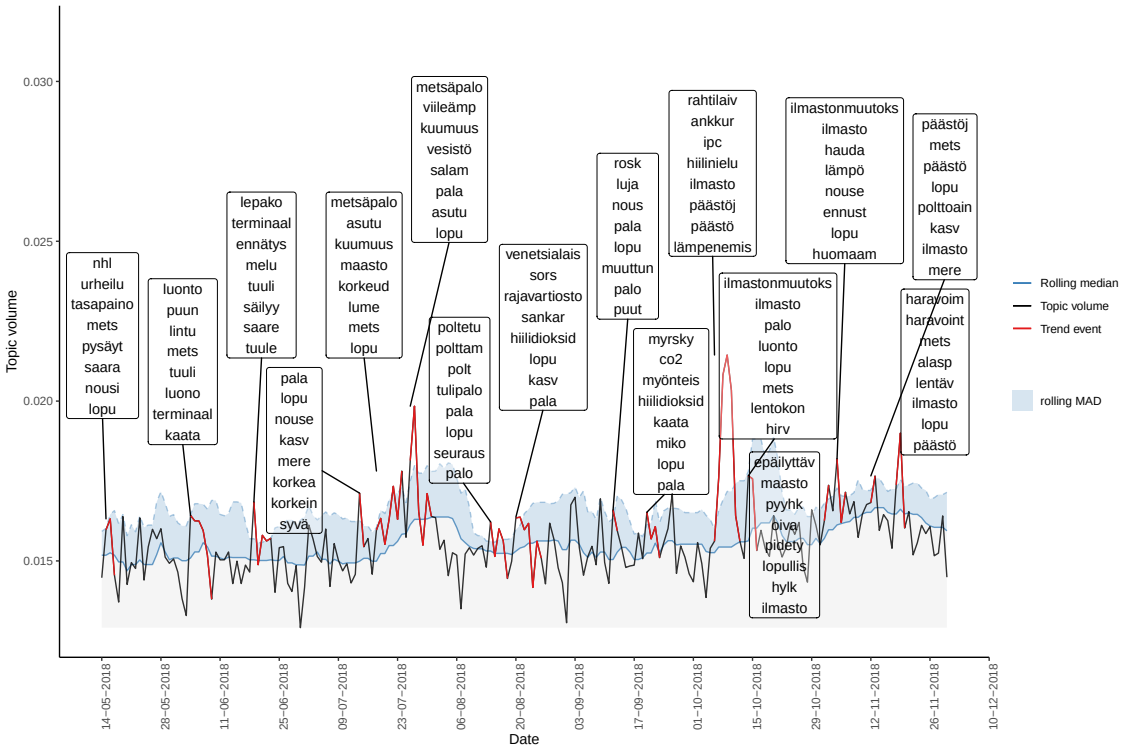


Figure A3: Suomi24 LDA 60, topic "Climate change".

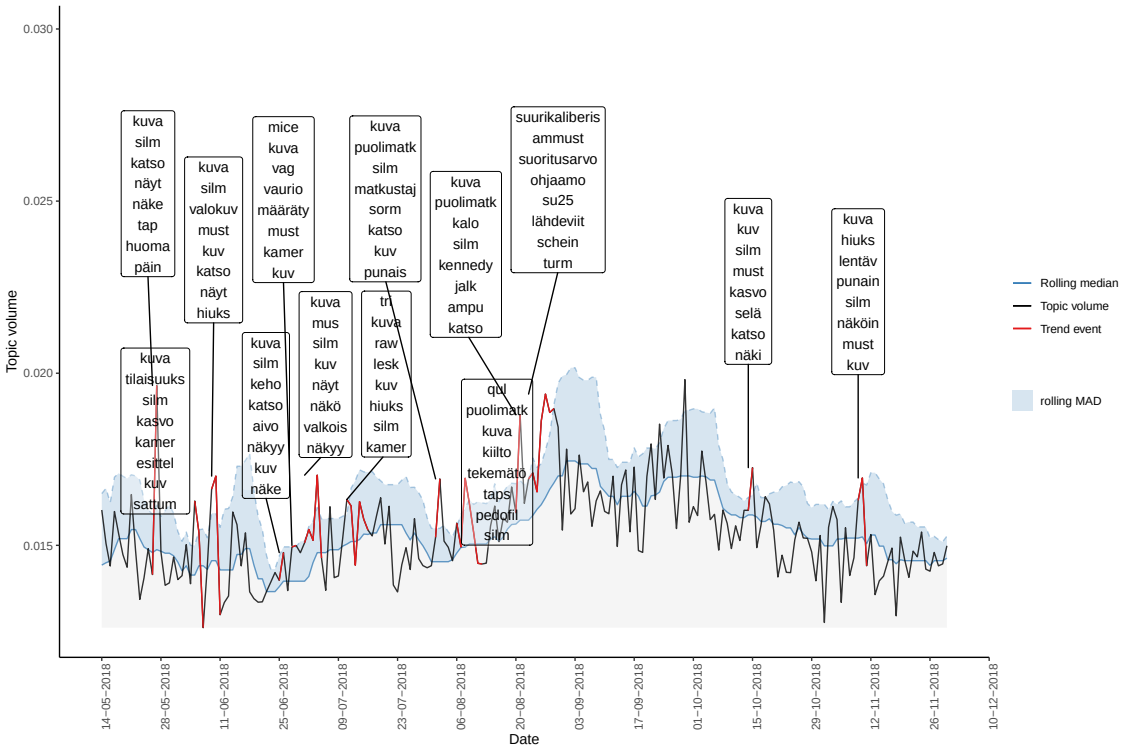


Figure A4: Suomi24 LDA 60, topic "Conspiracy theories".

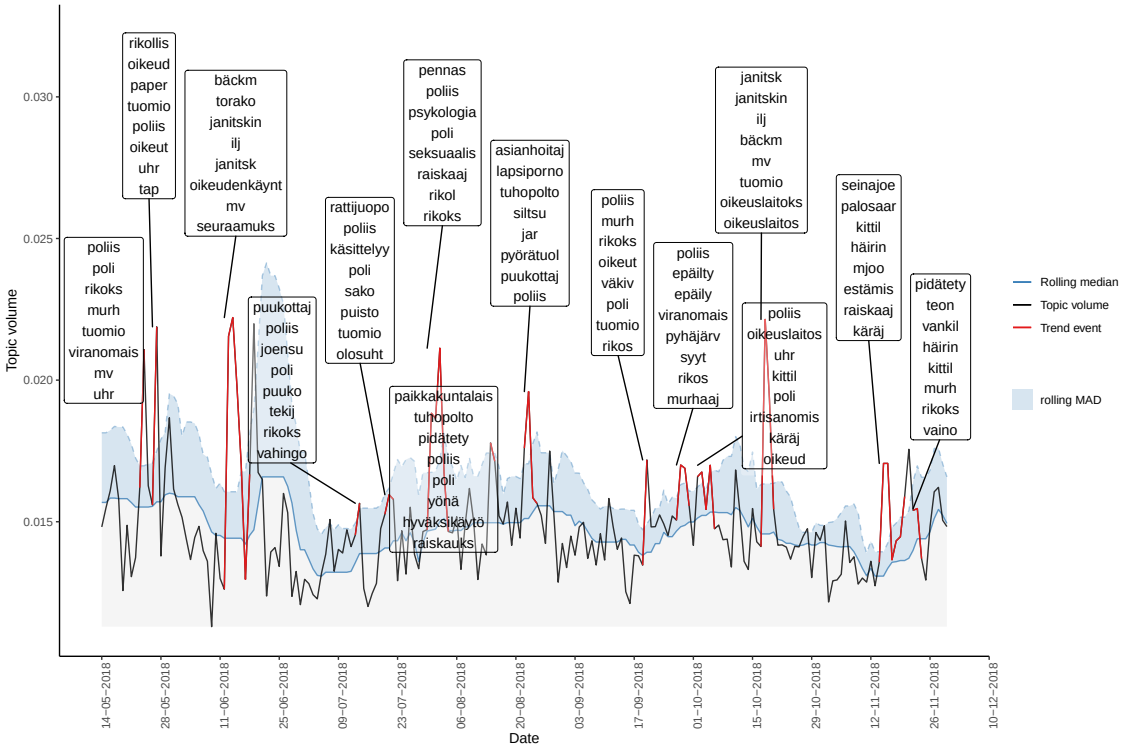


Figure A5: Suomi24 LDA 60, topic "Crimes".

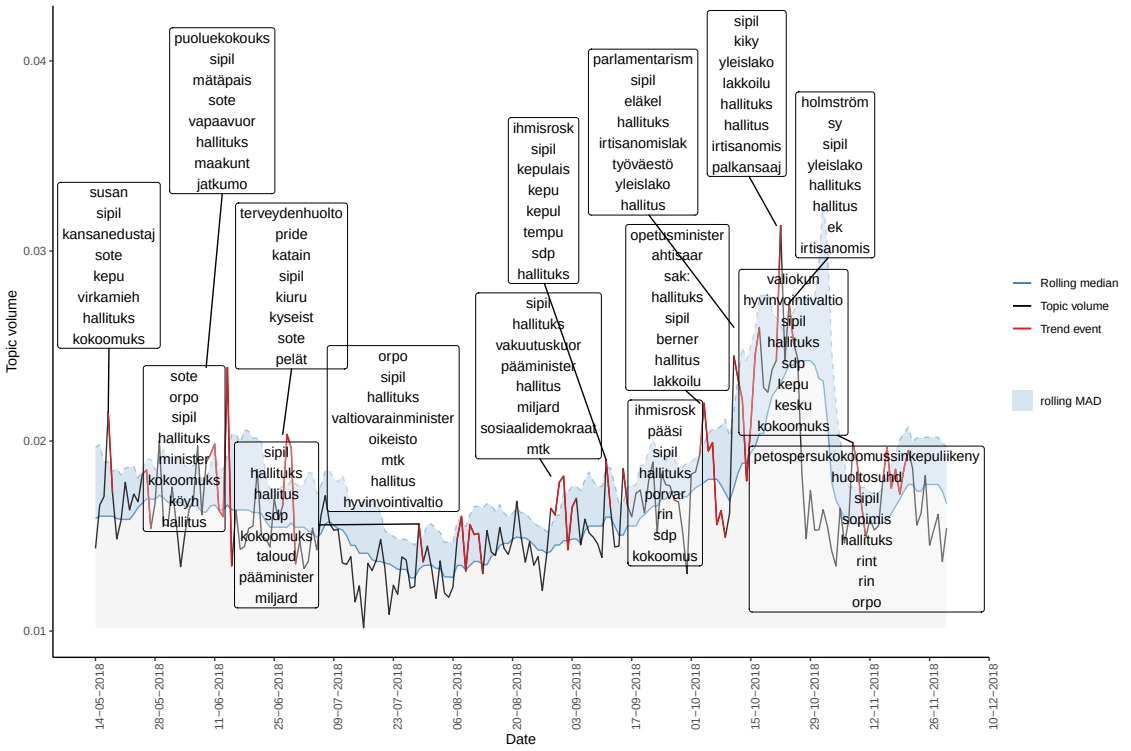


Figure A6: Suomi24 LDA 60, topic "Domestic Politics 1".

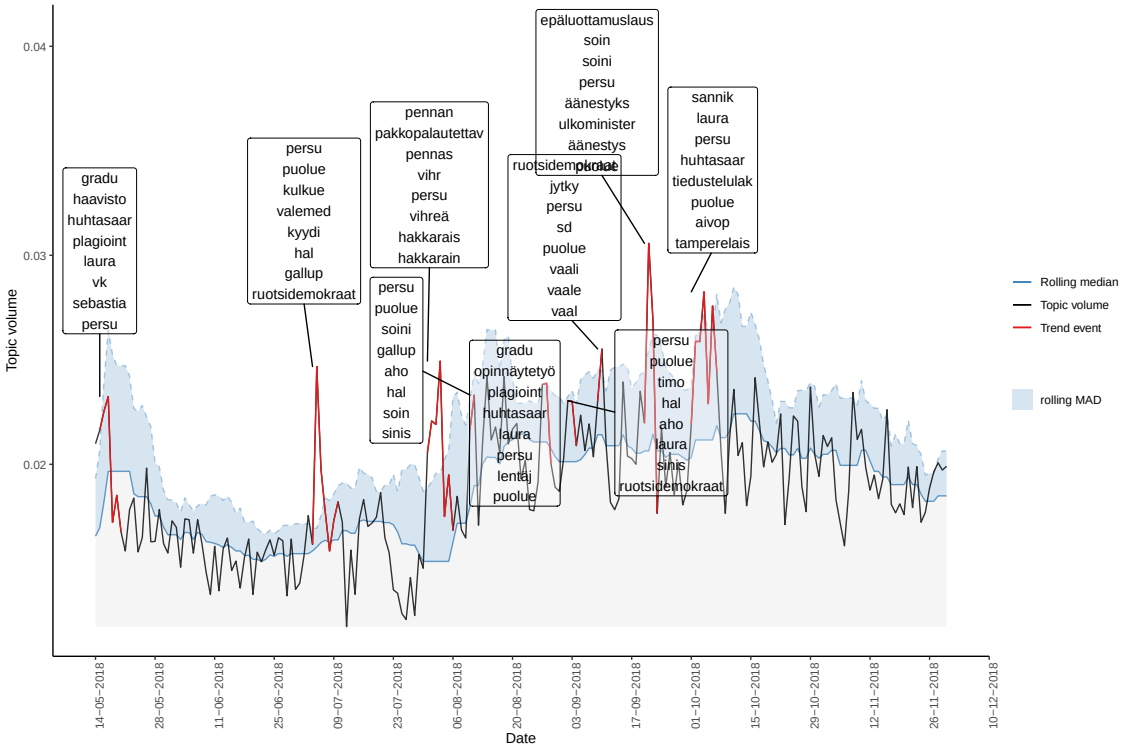


Figure A7: Suomi24 LDA 60, topic "Dometic Politics 2".

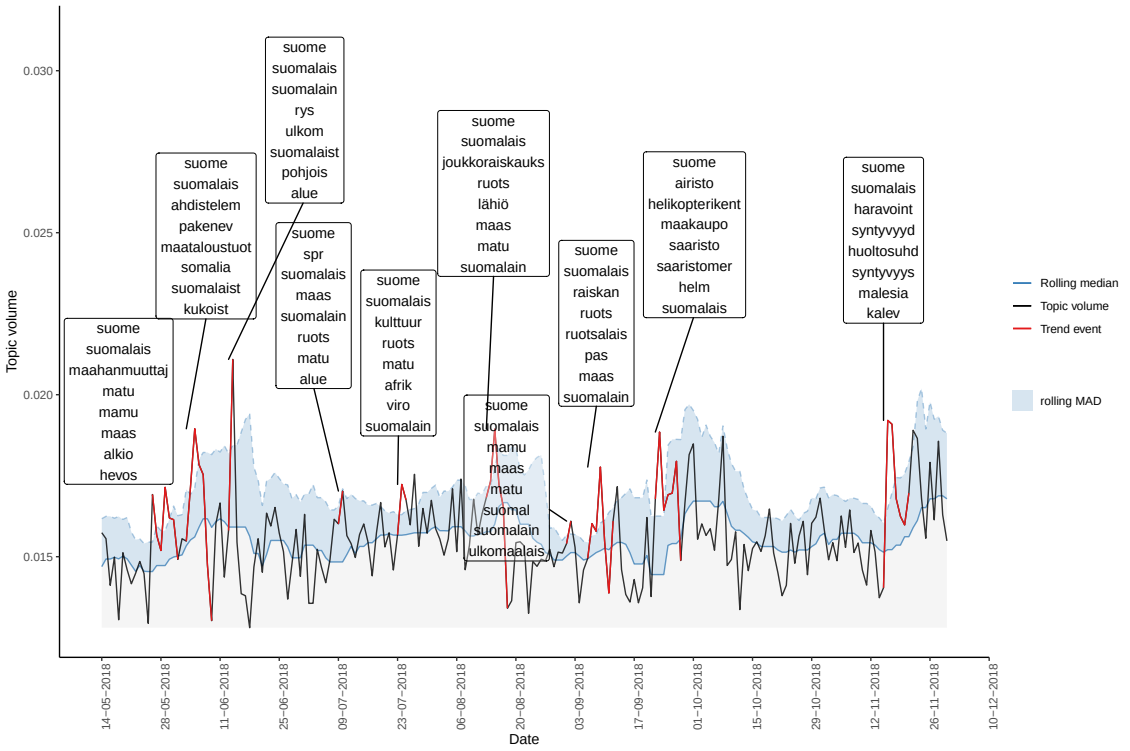


Figure A8: Suomi24 LDA 60, topic "Domestic Politics 3".

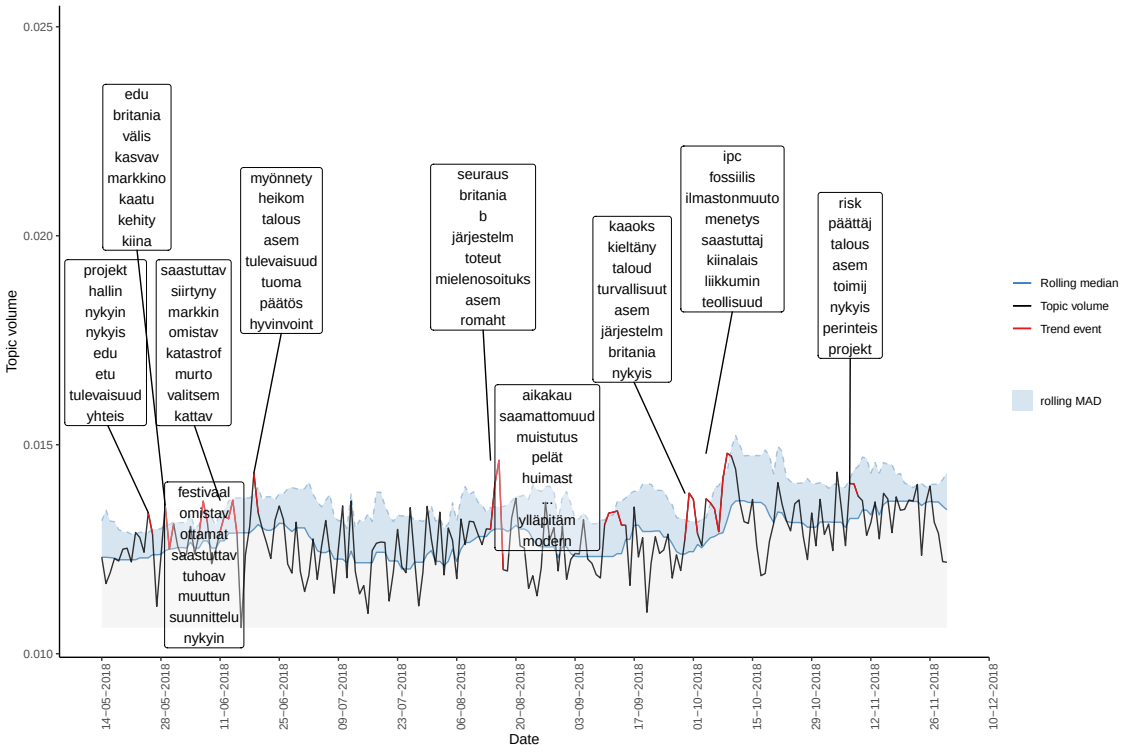


Figure A9: Suomi24 LDA 60, topic "Economics".

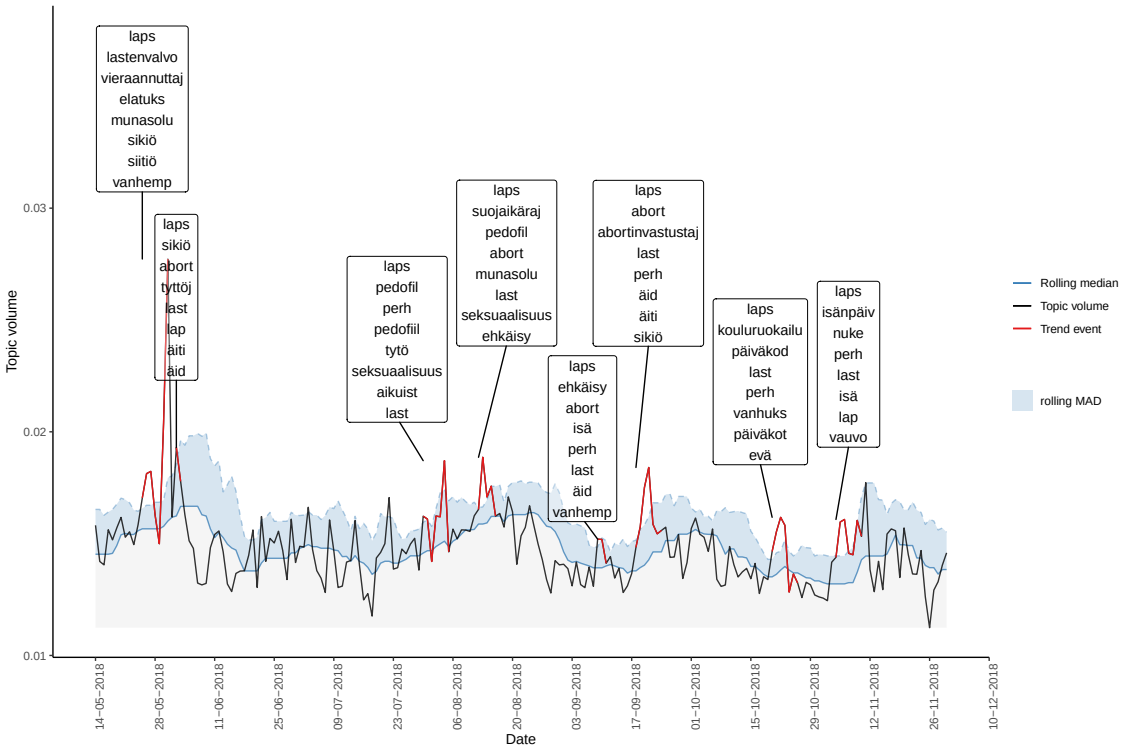


Figure A10: Suomi24 LDA 60, topic "Families".

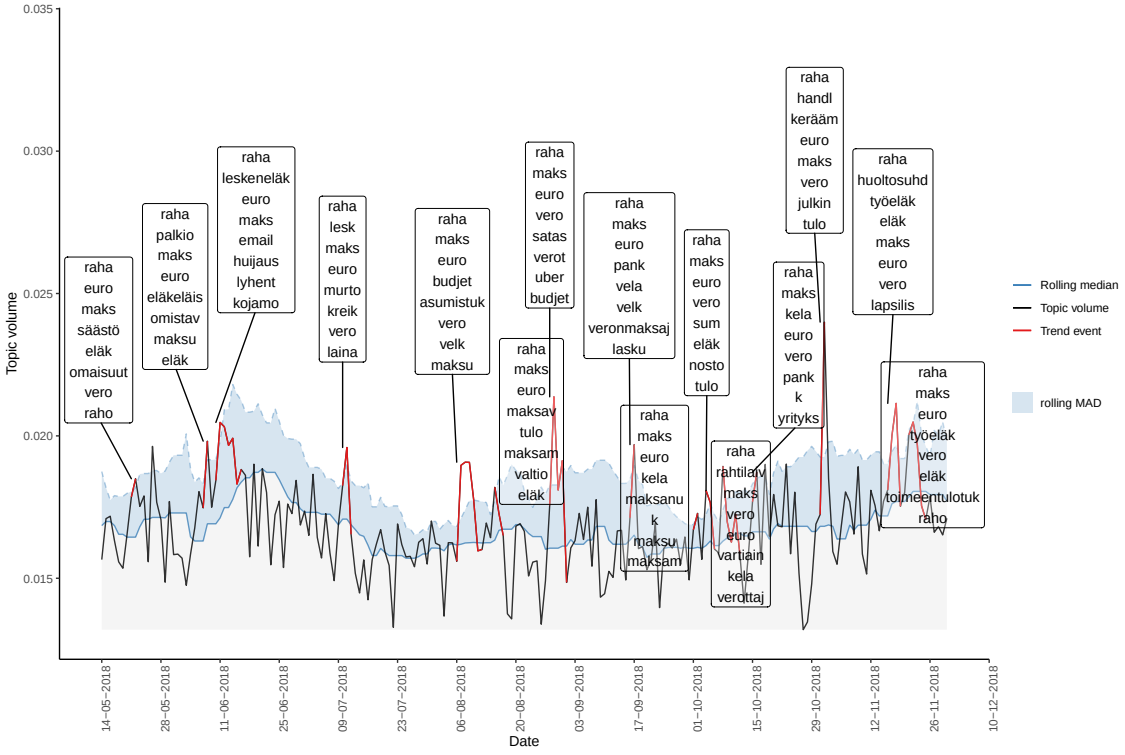


Figure A11: Suomi24 LDA 60, topic "Financial issues".

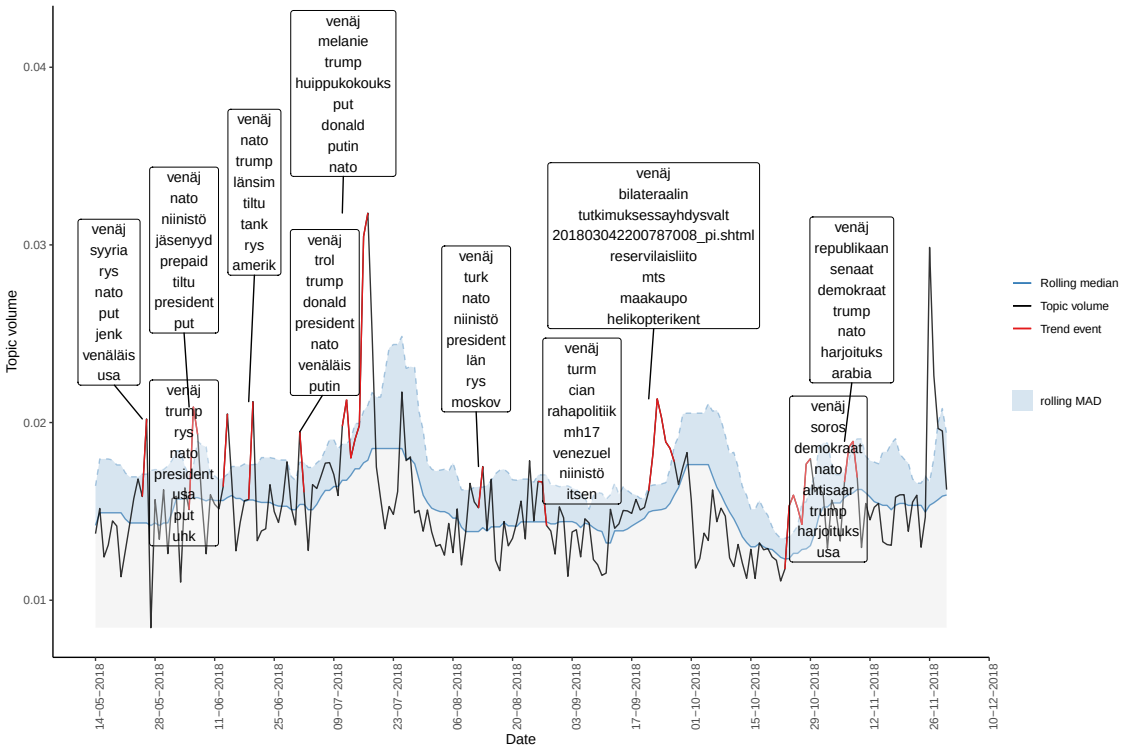


Figure A12: Suomi24 LDA 60, topic "Foreign Politics".

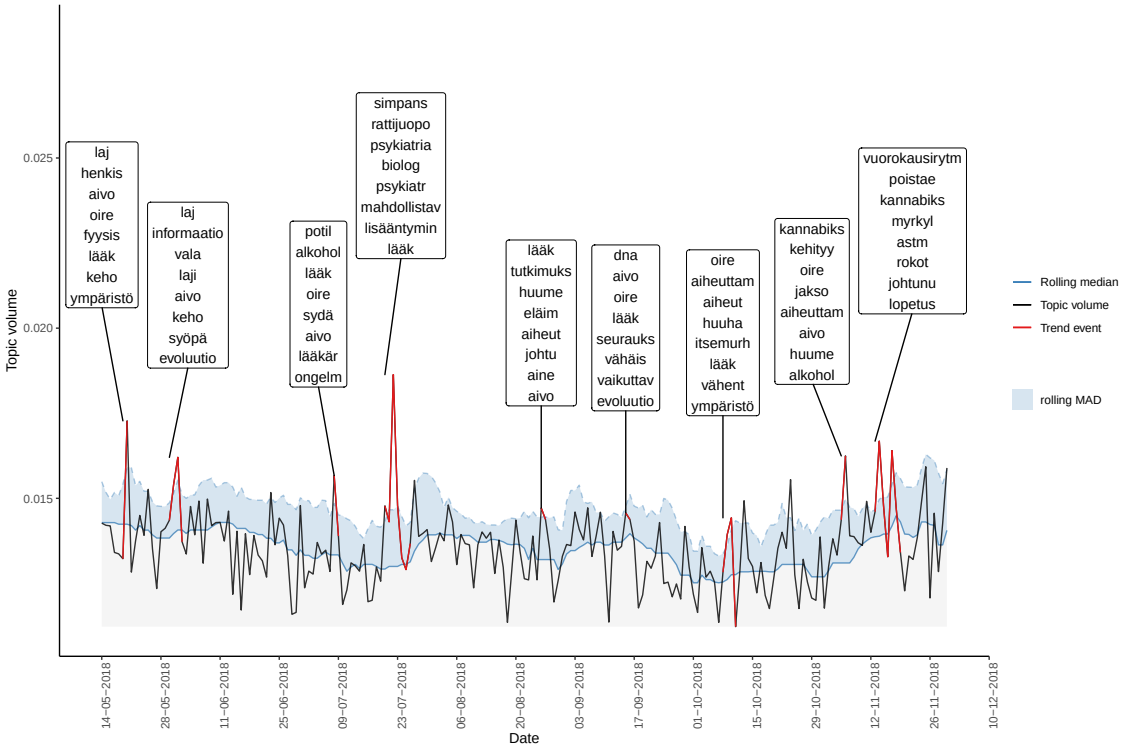


Figure A13: Suomi24 LDA 60, topic "Healthcare 1".

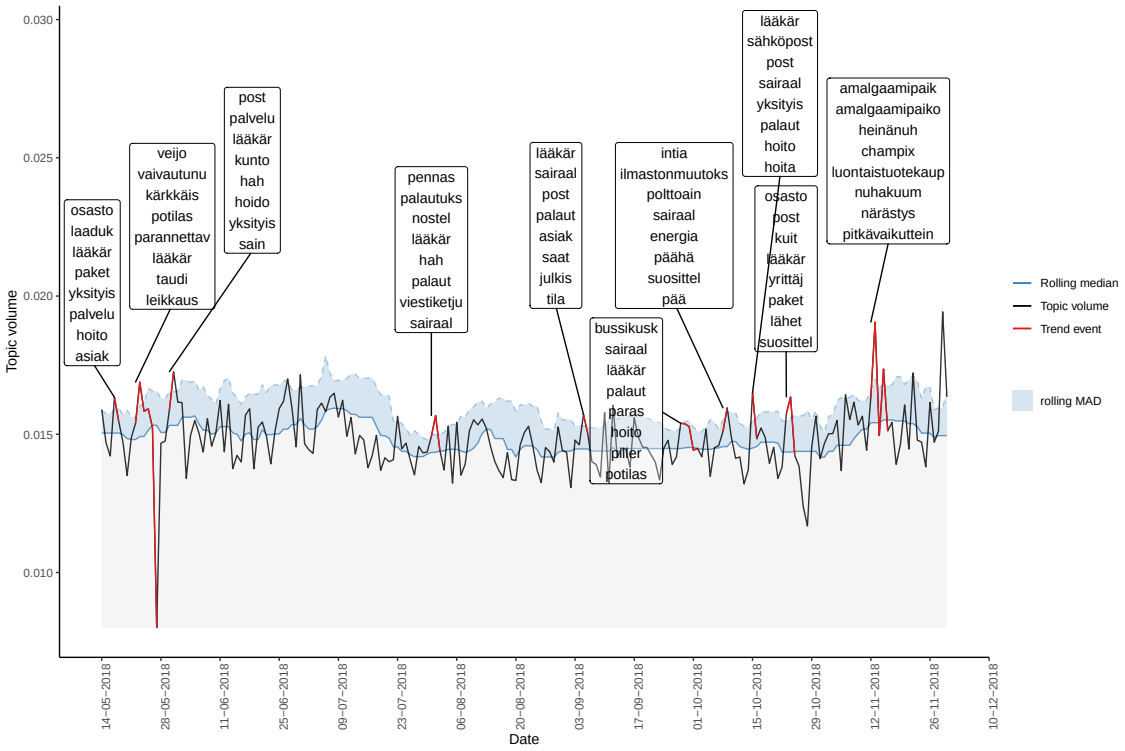


Figure A14: Suomi24 LDA 60, topic "Healthcare 2".

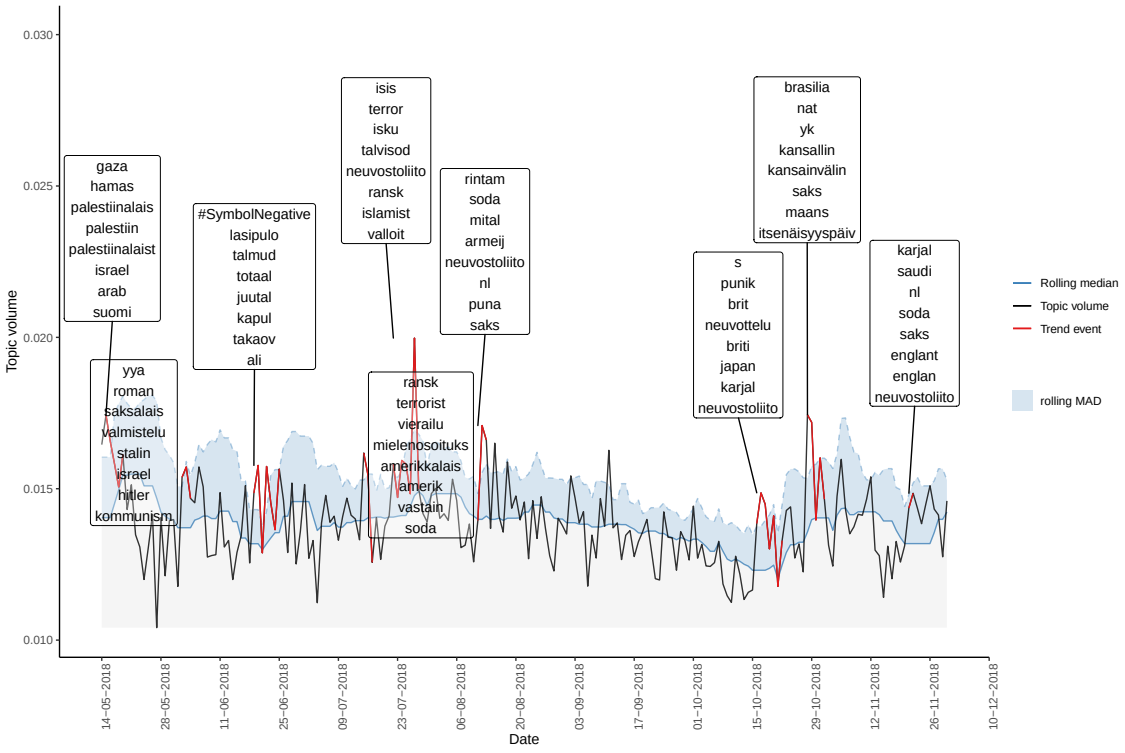


Figure A15: Suomi24 LDA 60, topic "Historical wars".

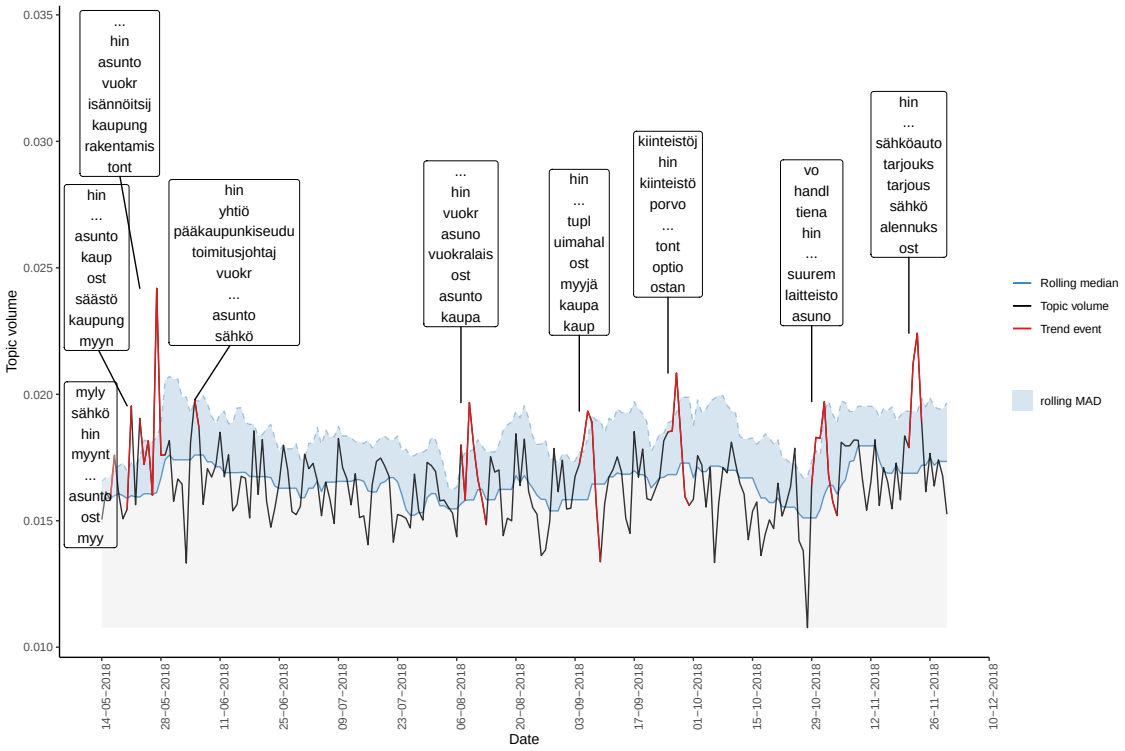


Figure A16: Suomi24 LDA 60, topic "Houses".

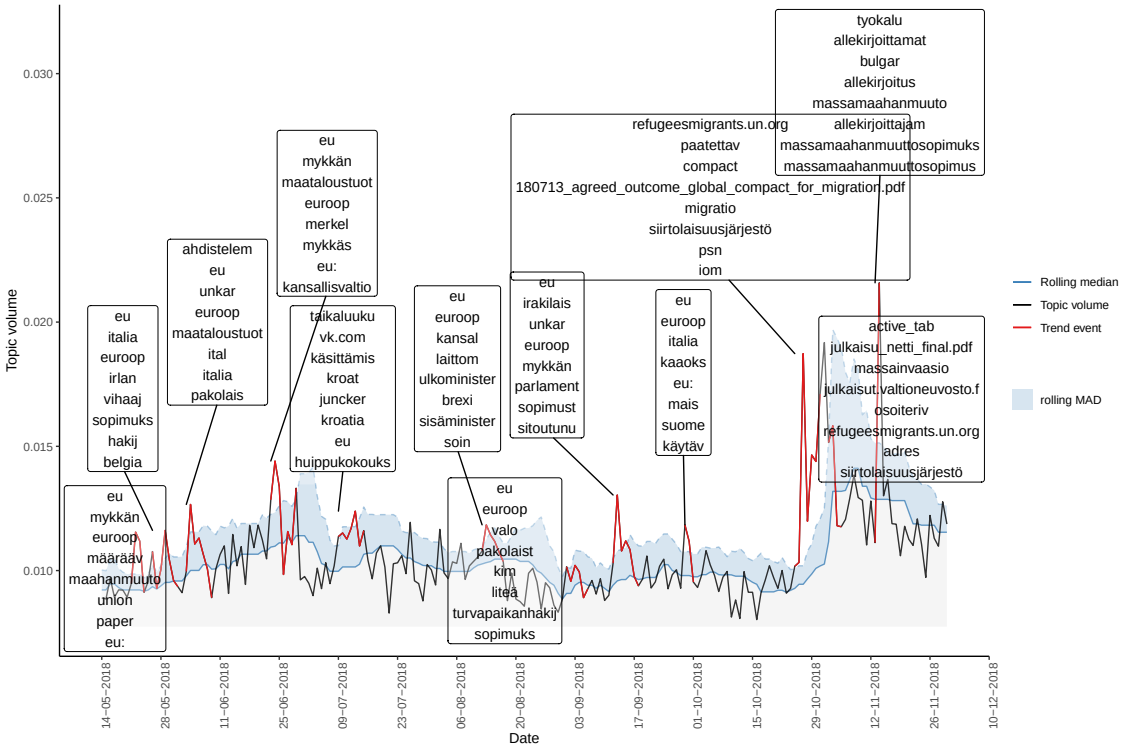


Figure A17: Suomi24 LDA 60, topic "Immigration 1".

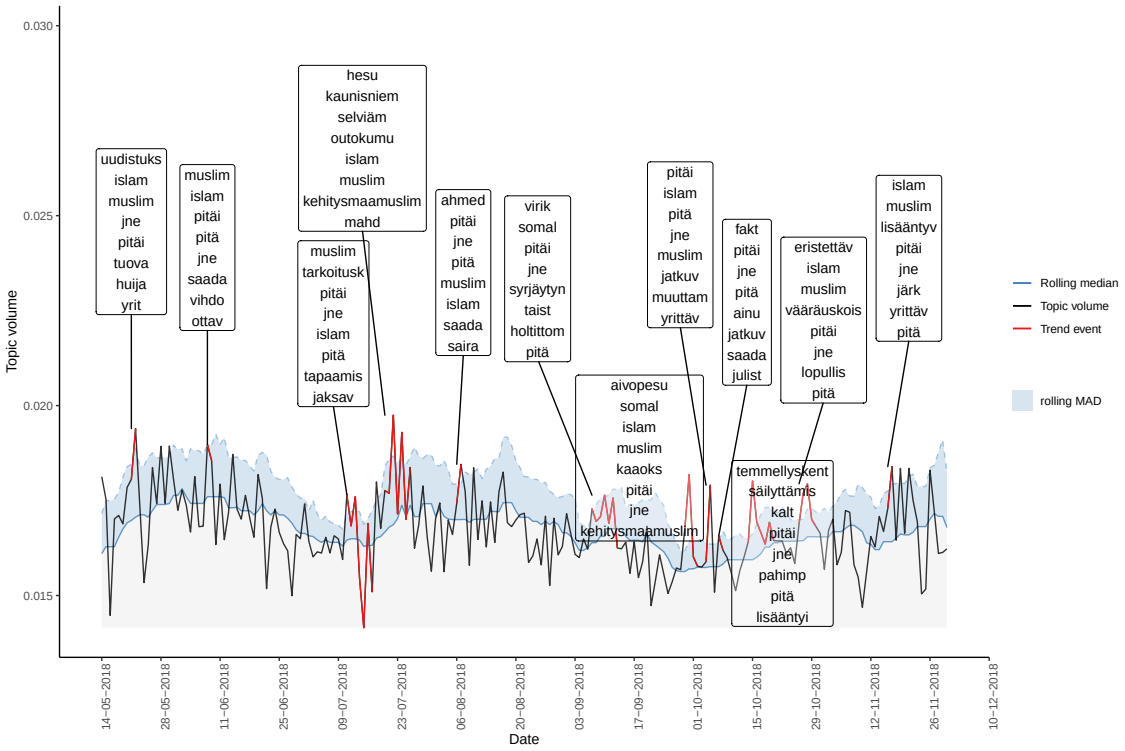


Figure A18: Suomi24 LDA 60, topic "Immigration 2".

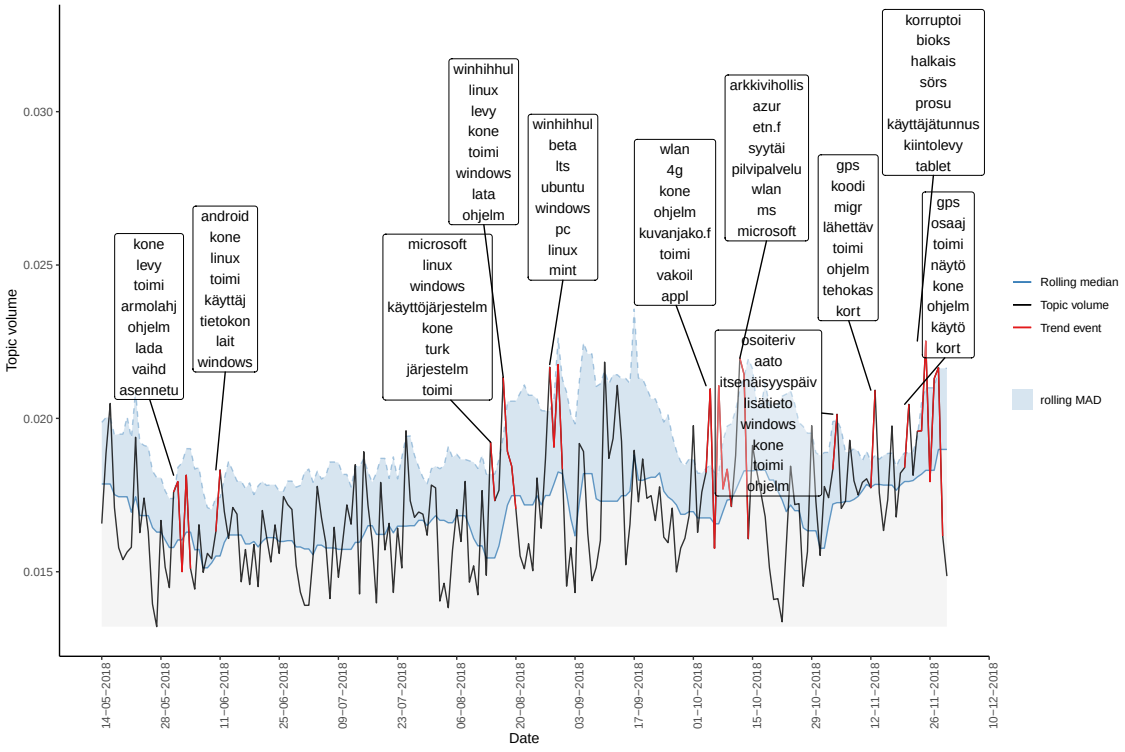


Figure A19: Suomi24 LDA 60, topic "Personal electronics".

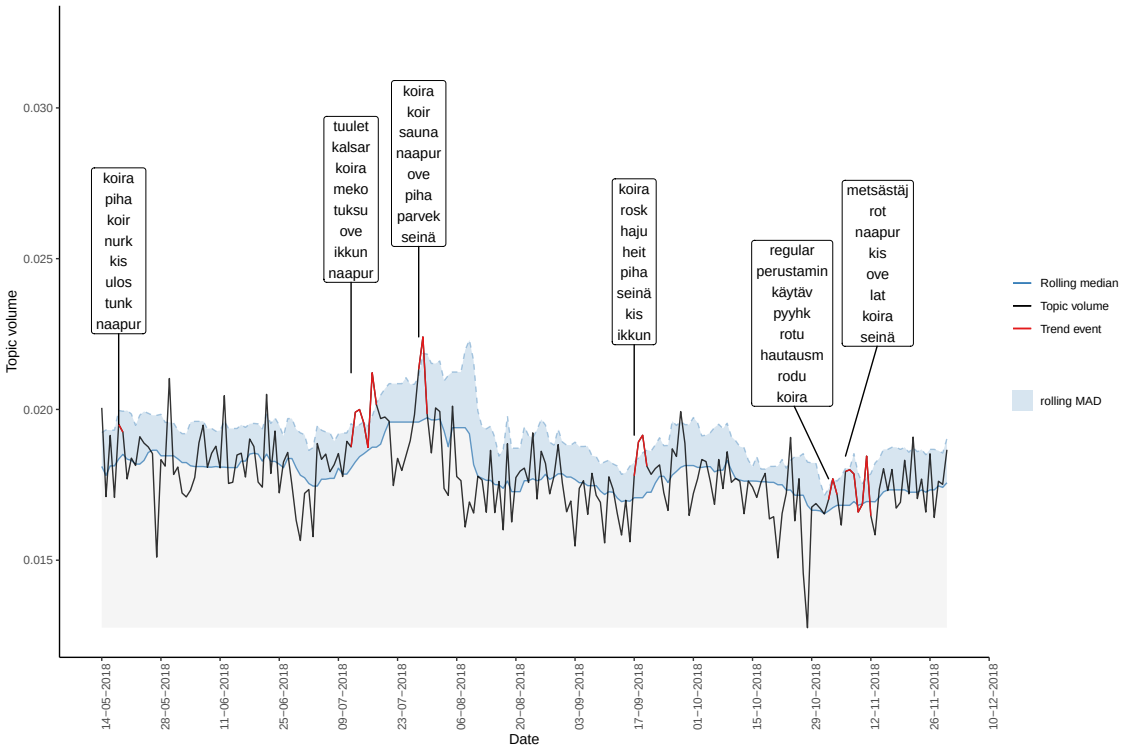


Figure A20: Suomi24 LDA 60, topic "Pets".

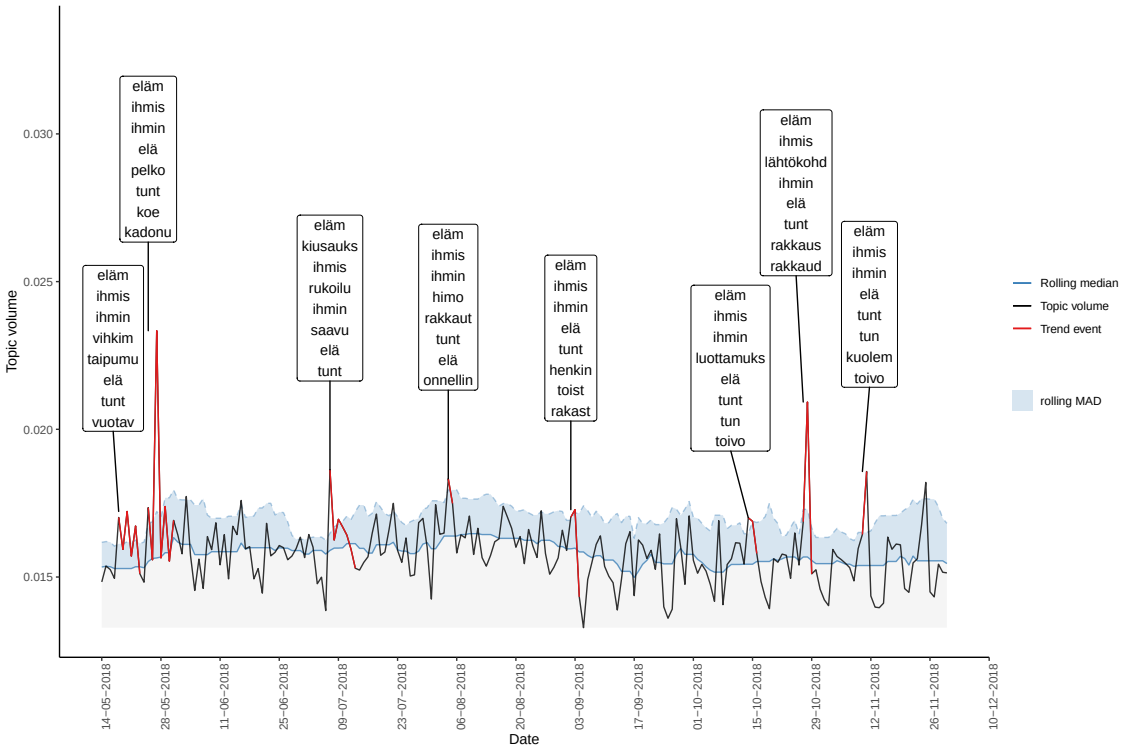


Figure A21: Suomi24 LDA 60, topic "Philosophical thoughts 1".

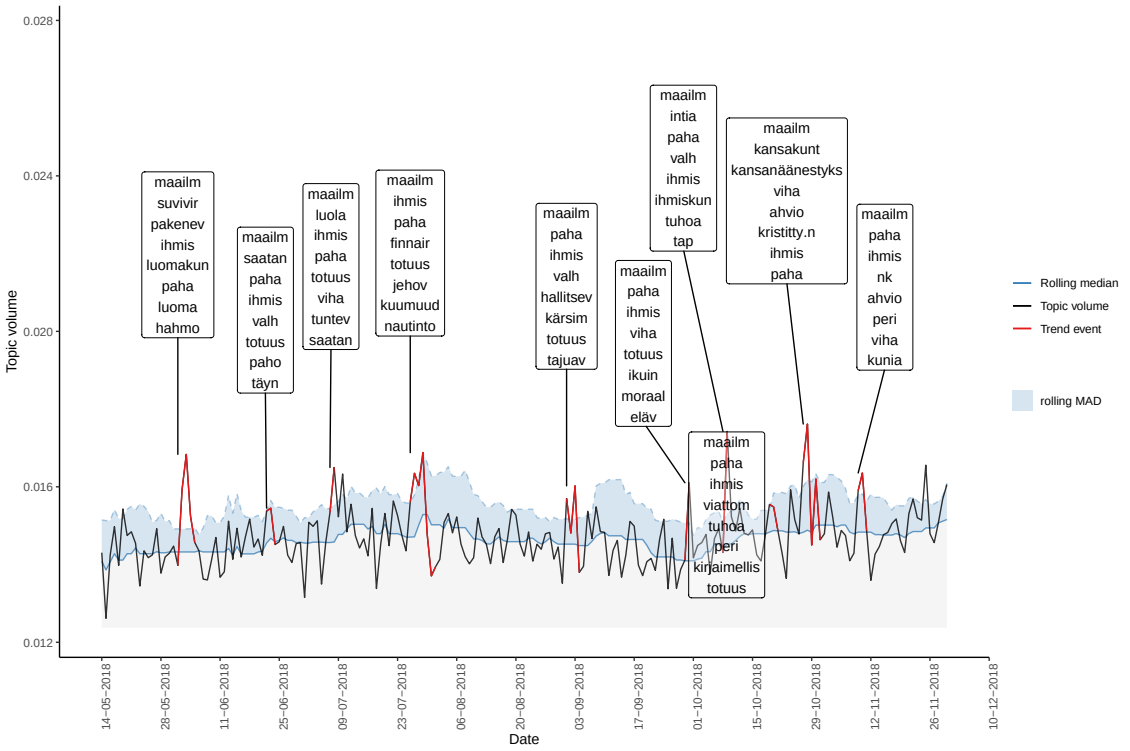


Figure A22: Suomi24 LDA 60, topic "Philosophical thoughts 2".

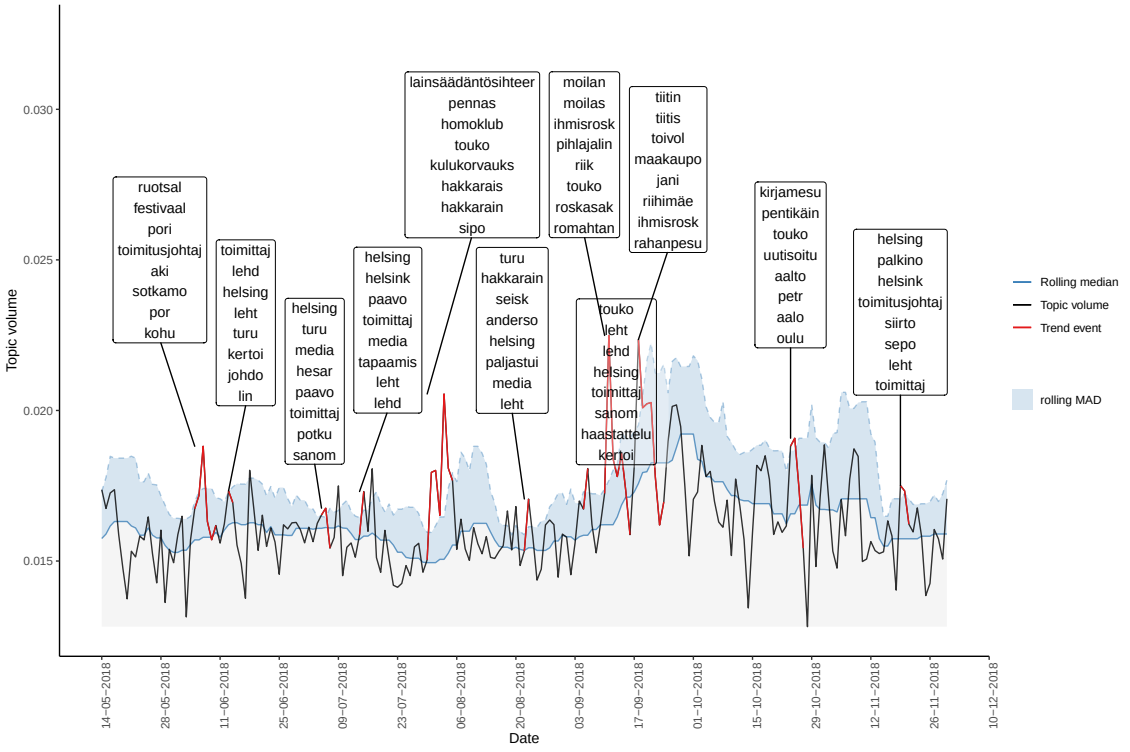


Figure A23: Suomi24 LDA 60, topic "Press media news".

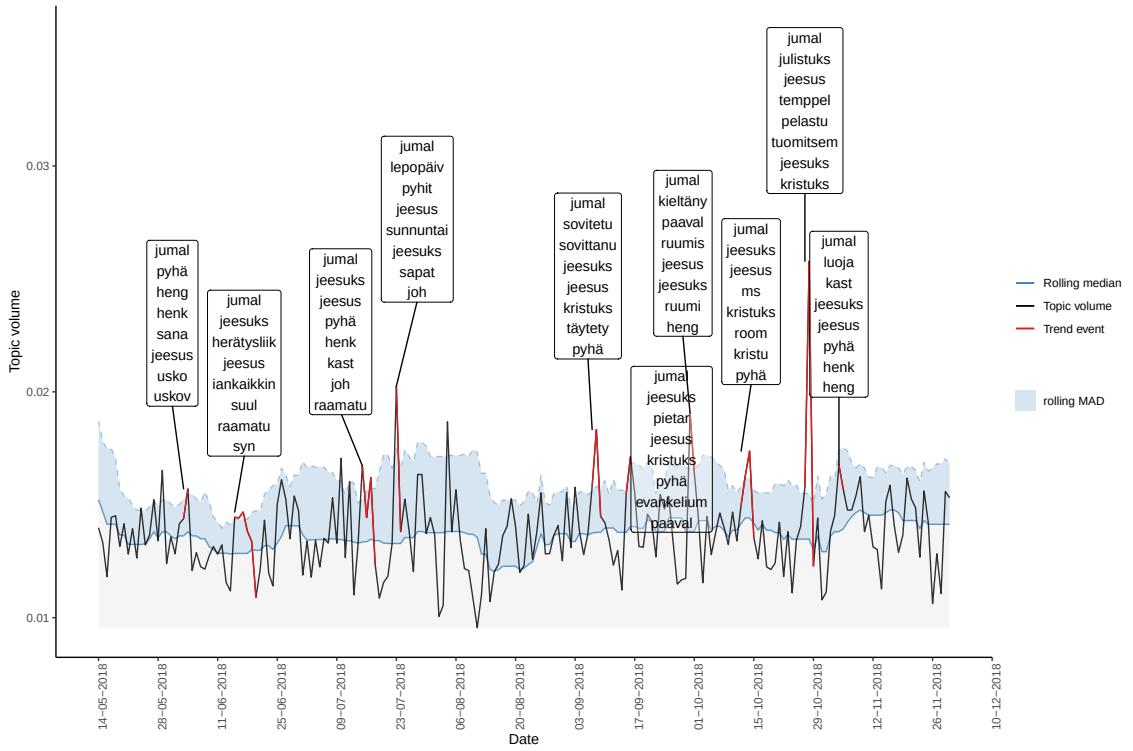


Figure A24: Suomi24 LDA 60, topic "Religion 1".

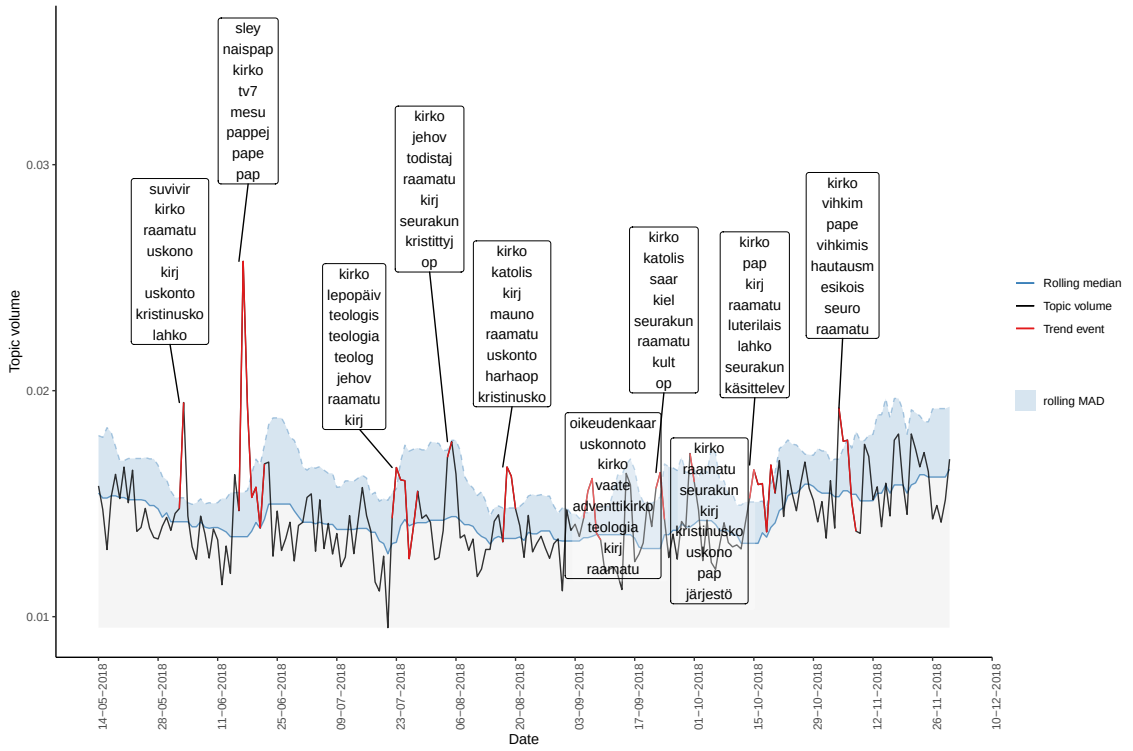


Figure A25: Suomi24 LDA 60, topic "Religion 2".

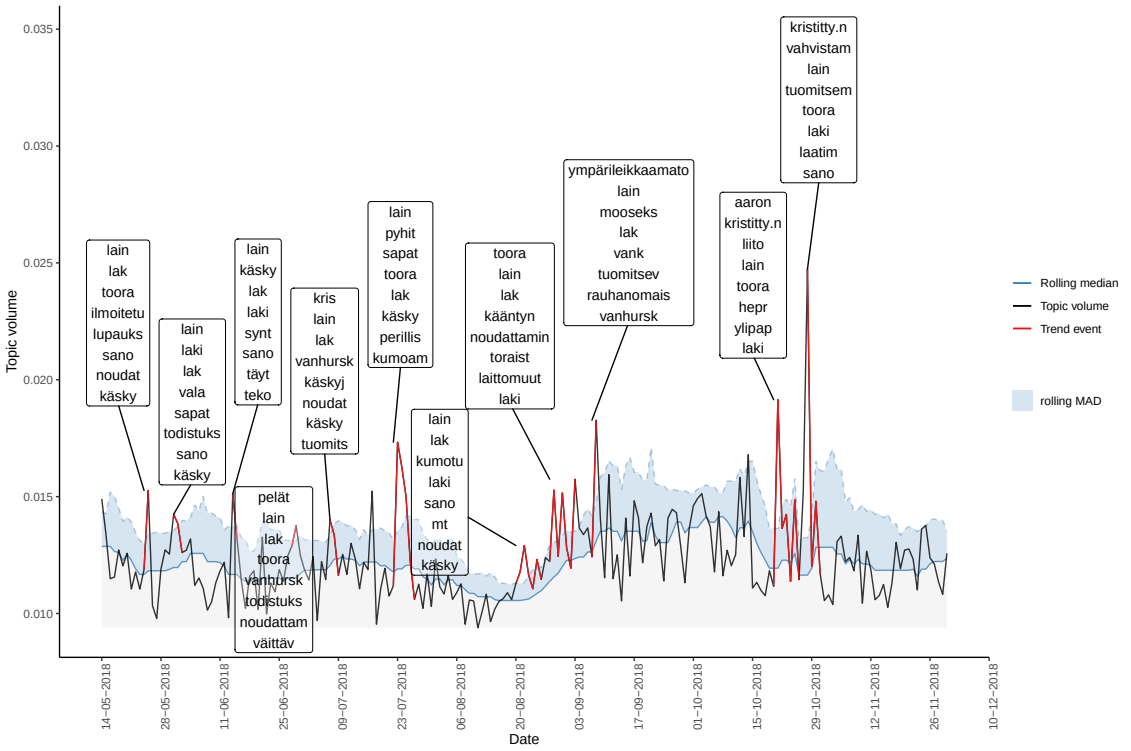


Figure A26: Suomi24 LDA 60, topic "Religion 3".

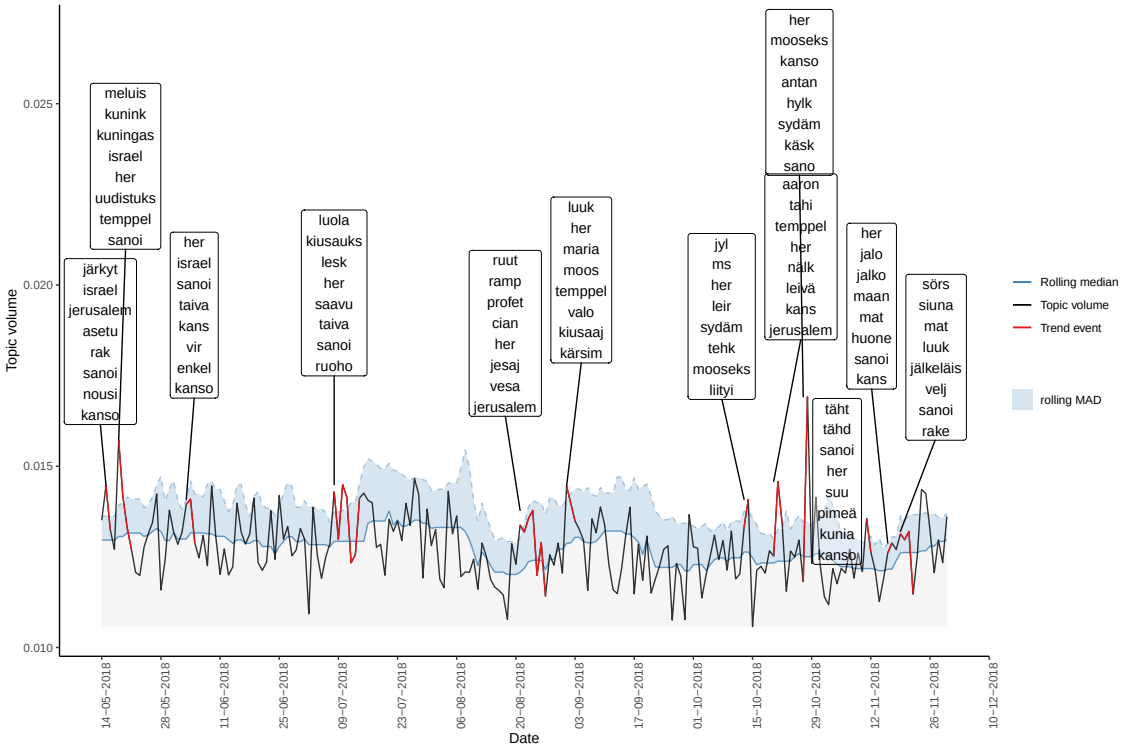


Figure A27: Suomi24 LDA 60, topic "Religion 4".

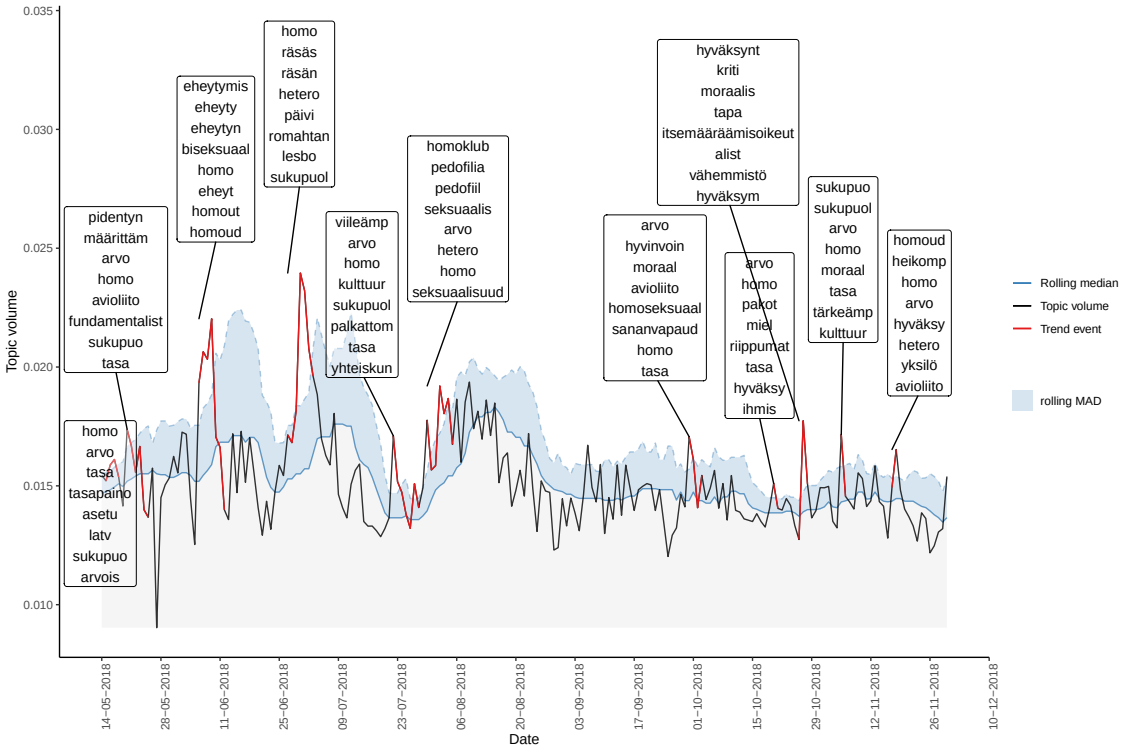


Figure A28: Suomi24 LDA 60, topic "Same sex relationships".

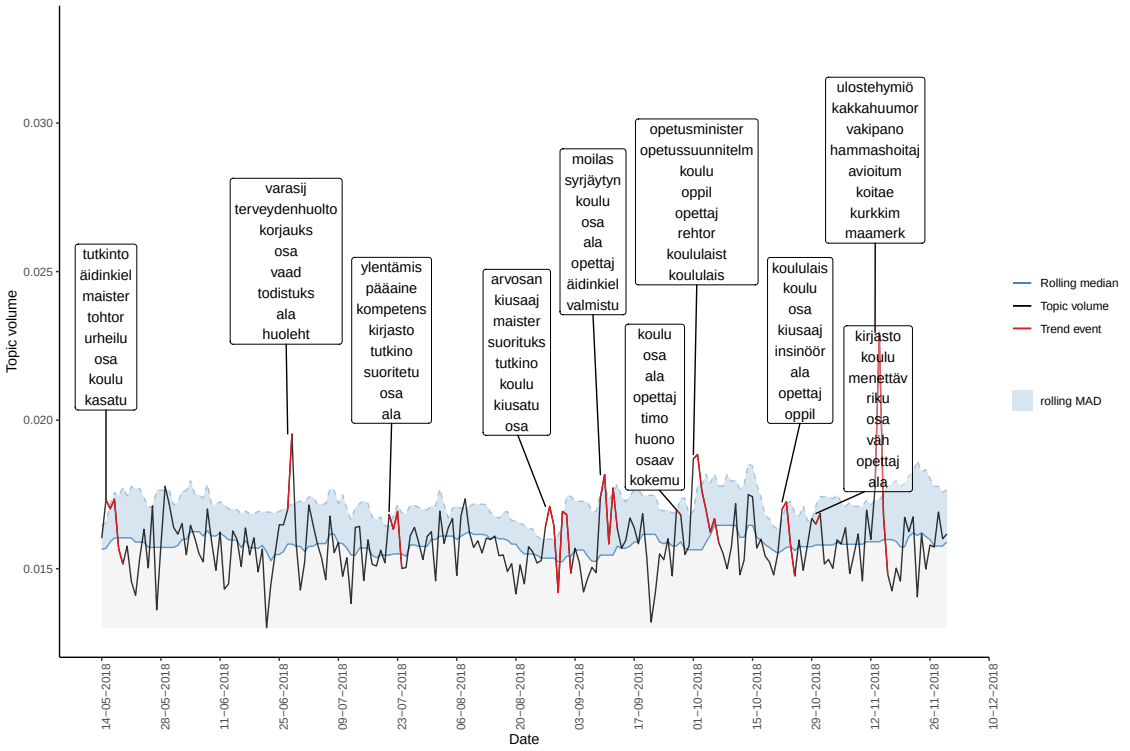


Figure A29: Suomi24 LDA 60, topic "Schools".

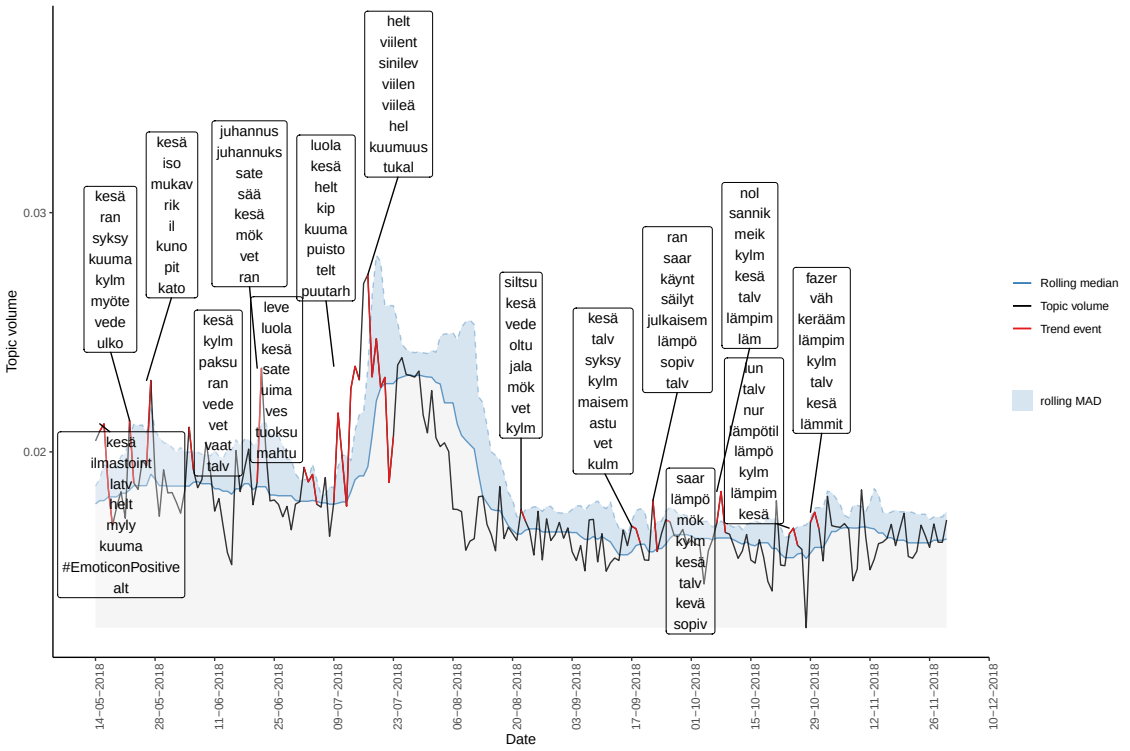


Figure A30: Suomi24 LDA 60, topic "Seasons".

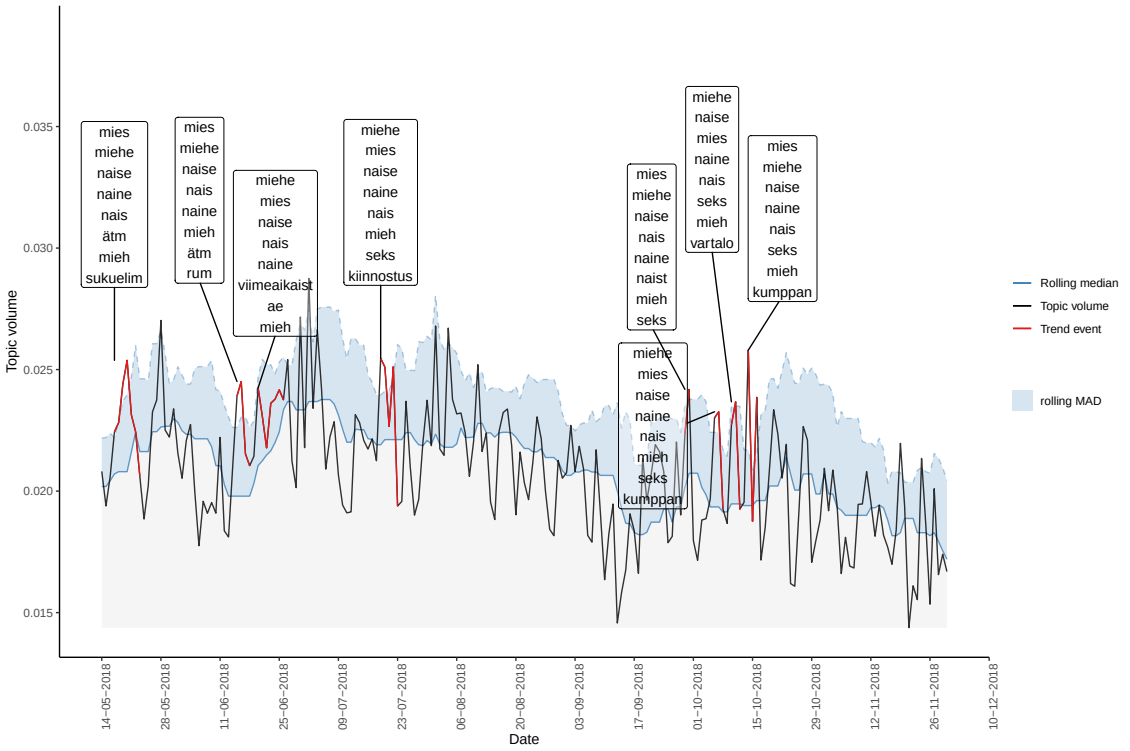


Figure A31: Suomi24 LDA 60, topic "Sexual relationships".

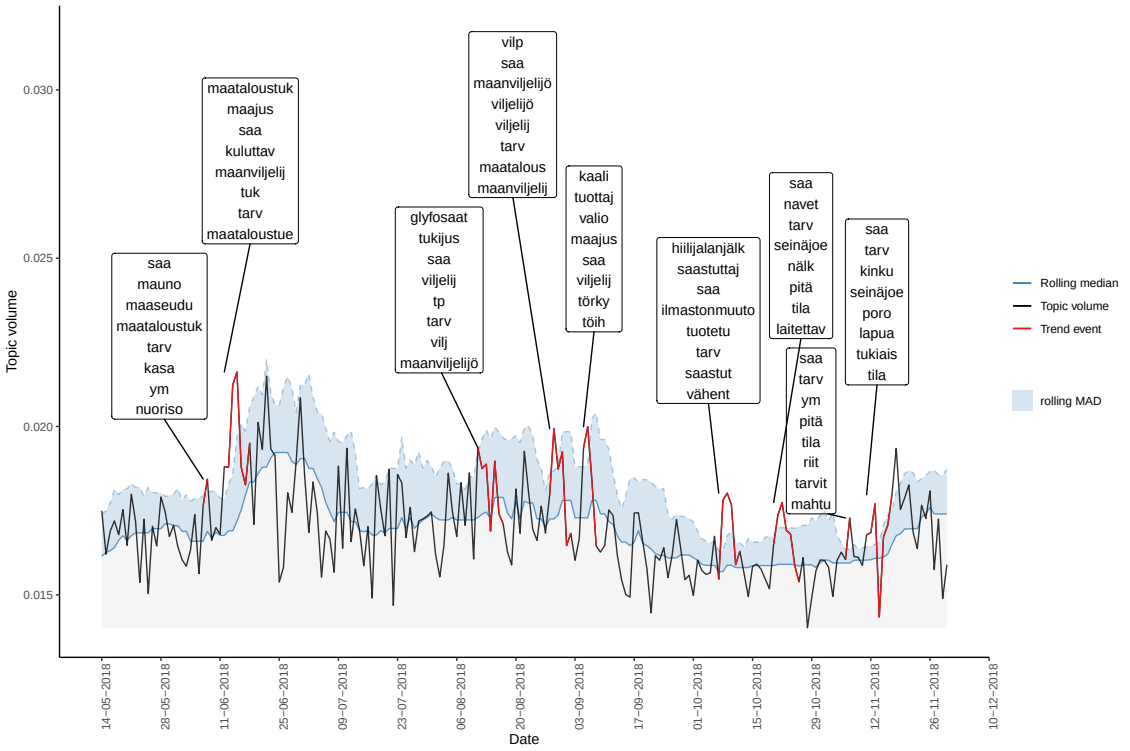


Figure A32: Suomi24 LDA 60, topic "Societal support".

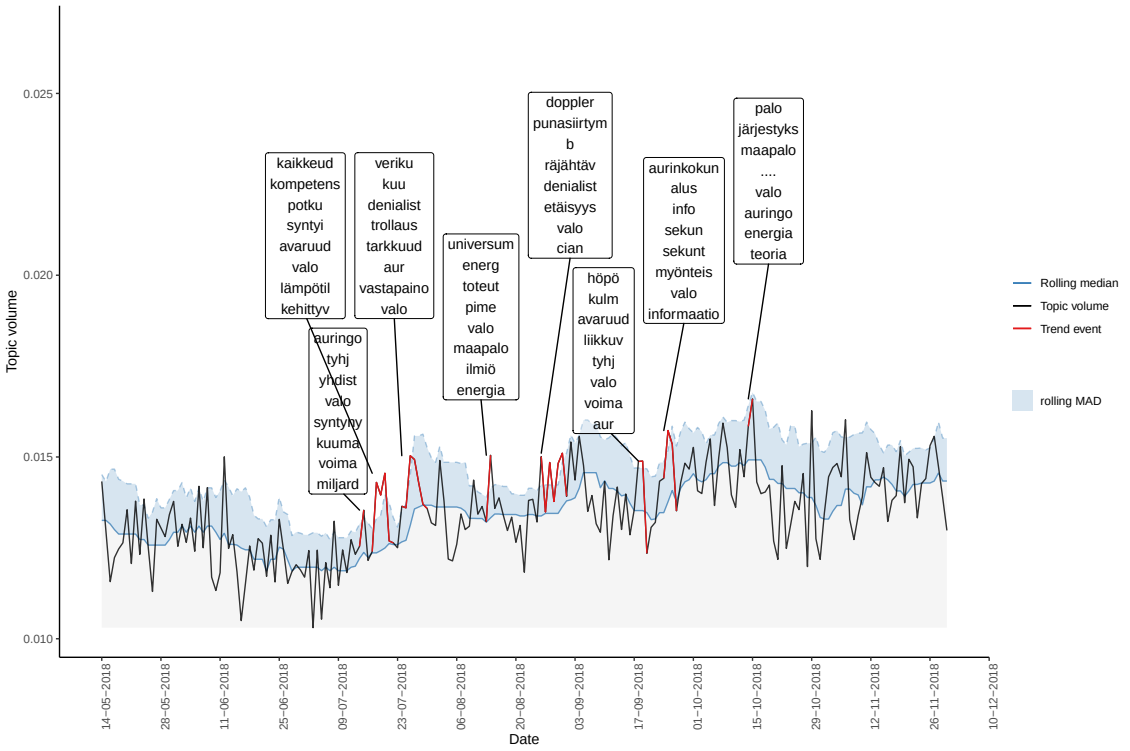


Figure A33: Suomi24 LDA 60, topic "The Universe".

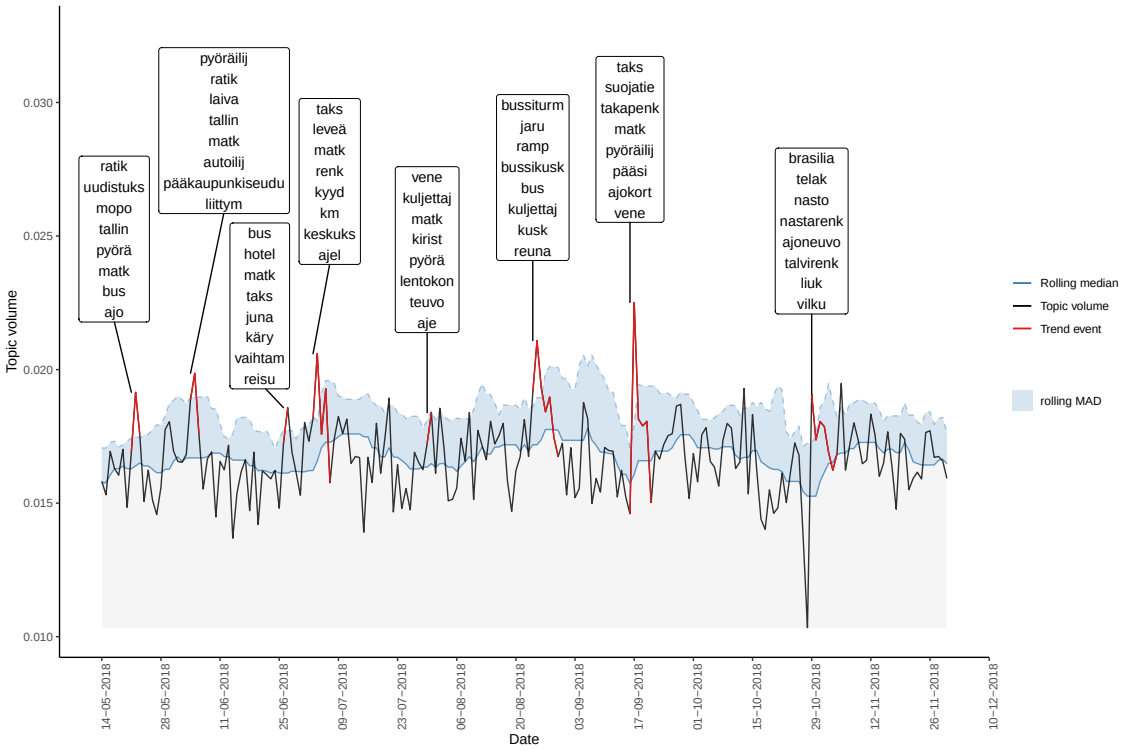


Figure A34: Suomi24 LDA 60, topic "Traffic".

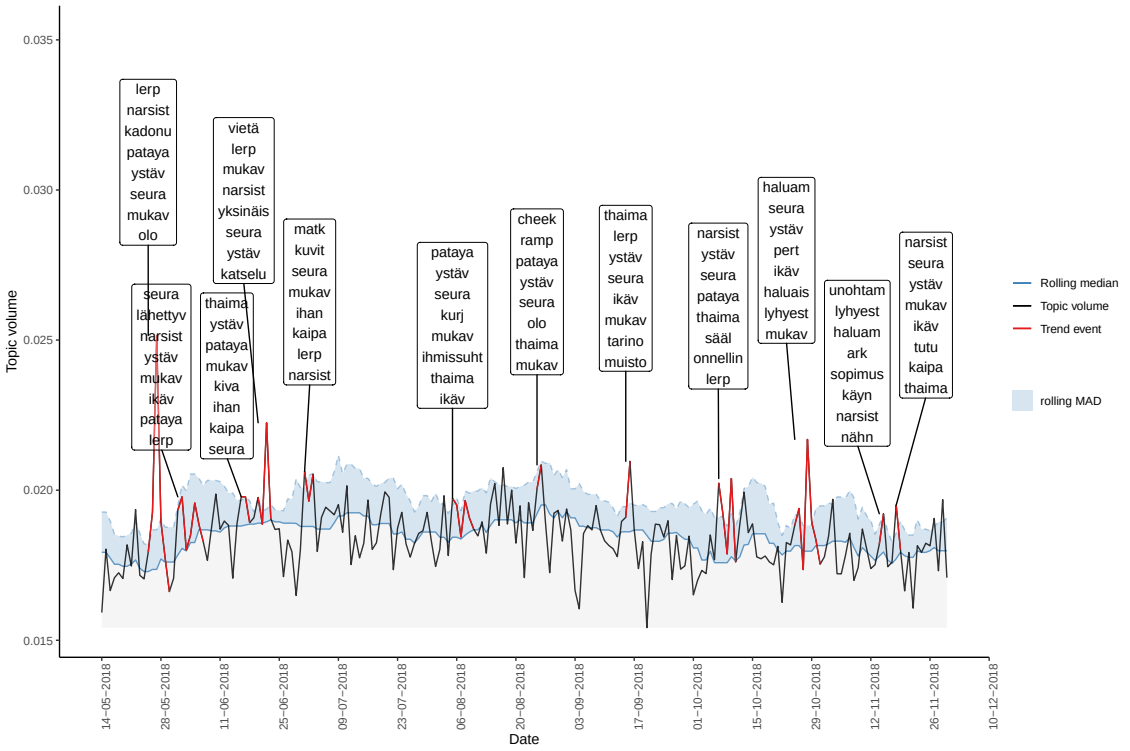


Figure A35: Suomi24 LDA 60, topic "Thailand travel and dating".

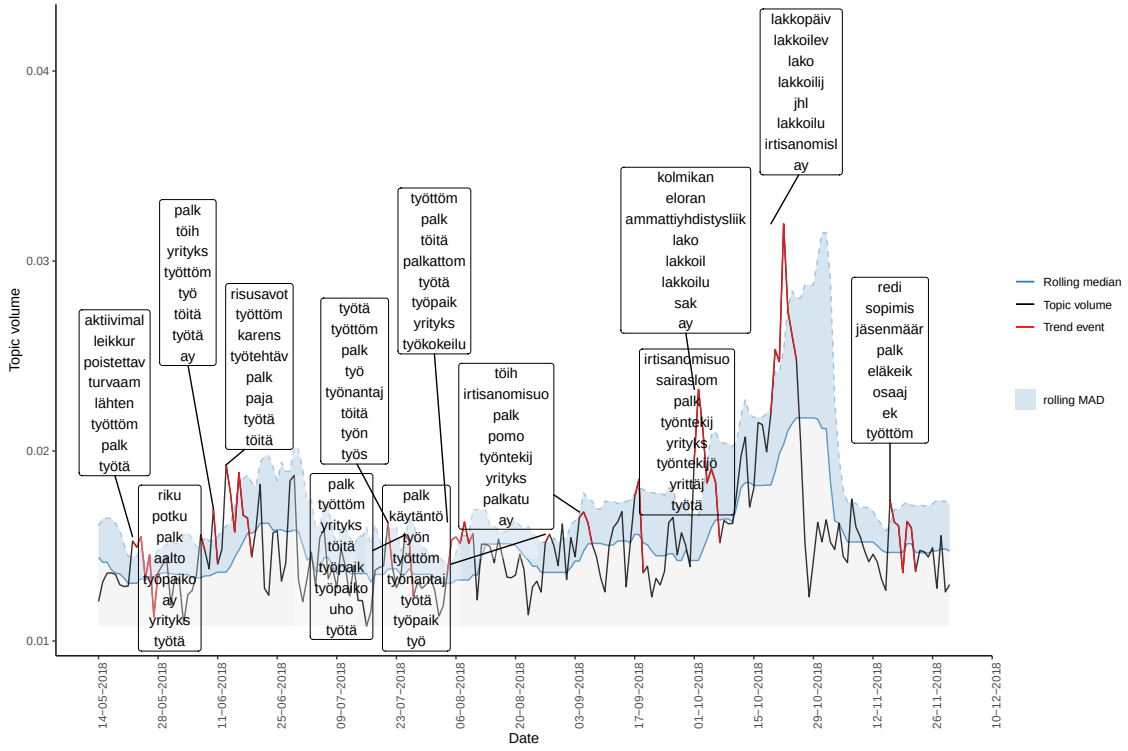


Figure A36: Suomi24 LDA 60, topic "Working".

B Trend events in `vauva.fi`

This appendix lists the figures of topical trend volumes and identified trend events for `vauva.fi` LDA model with 40 topics. Only topics that were evaluated to have a coherent semantic meaning based on the top 10 terms by relevance, are listed. Refer to Figure 4.1.4 for the relative sizes and top words for each topic.

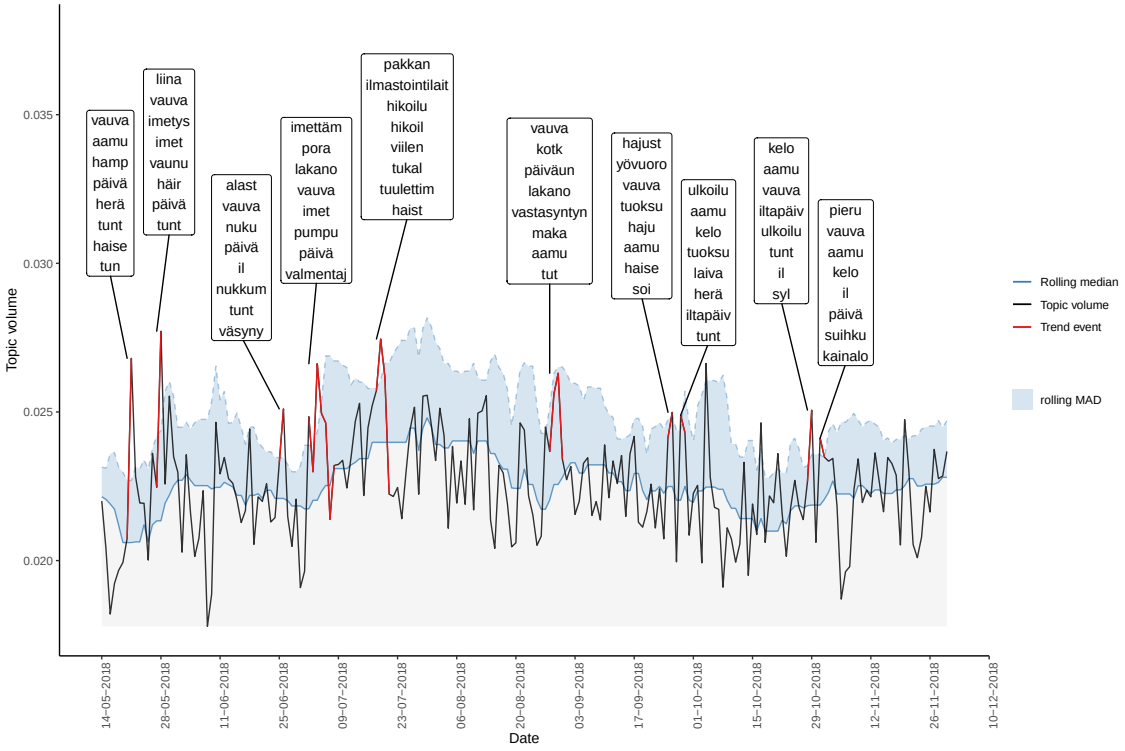


Figure B1: Vauva LDA 40, topic "Babies".

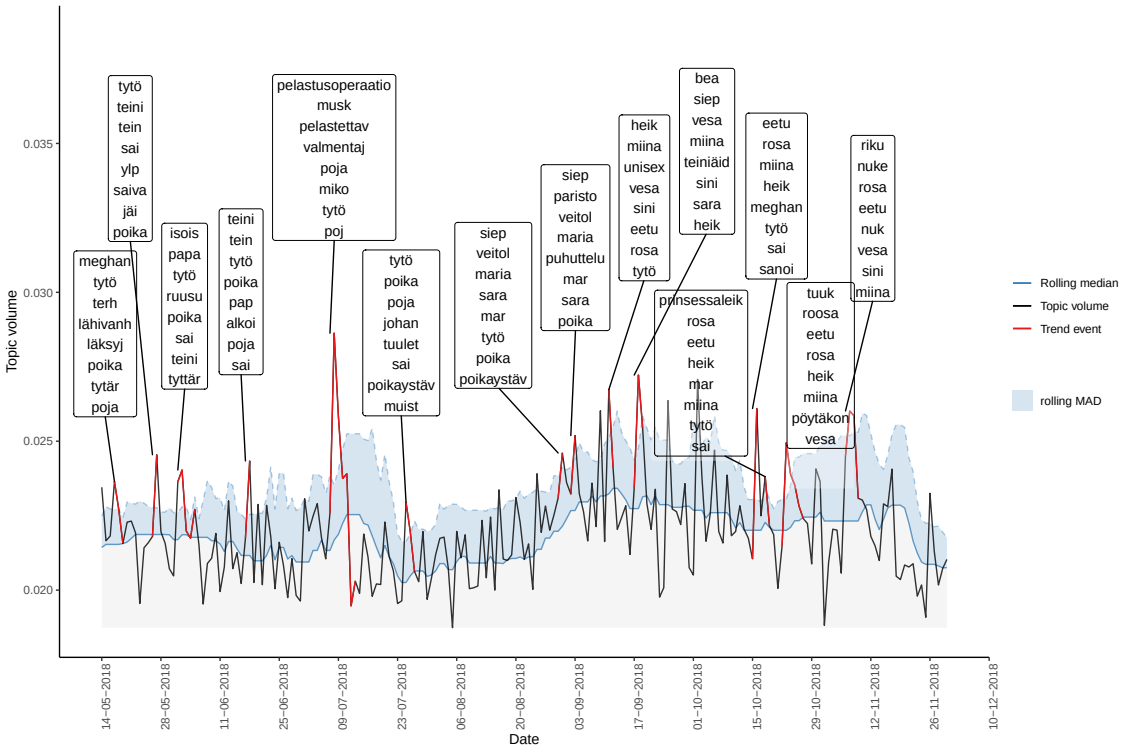


Figure B2: Vauva LDA 40, topic "Children".

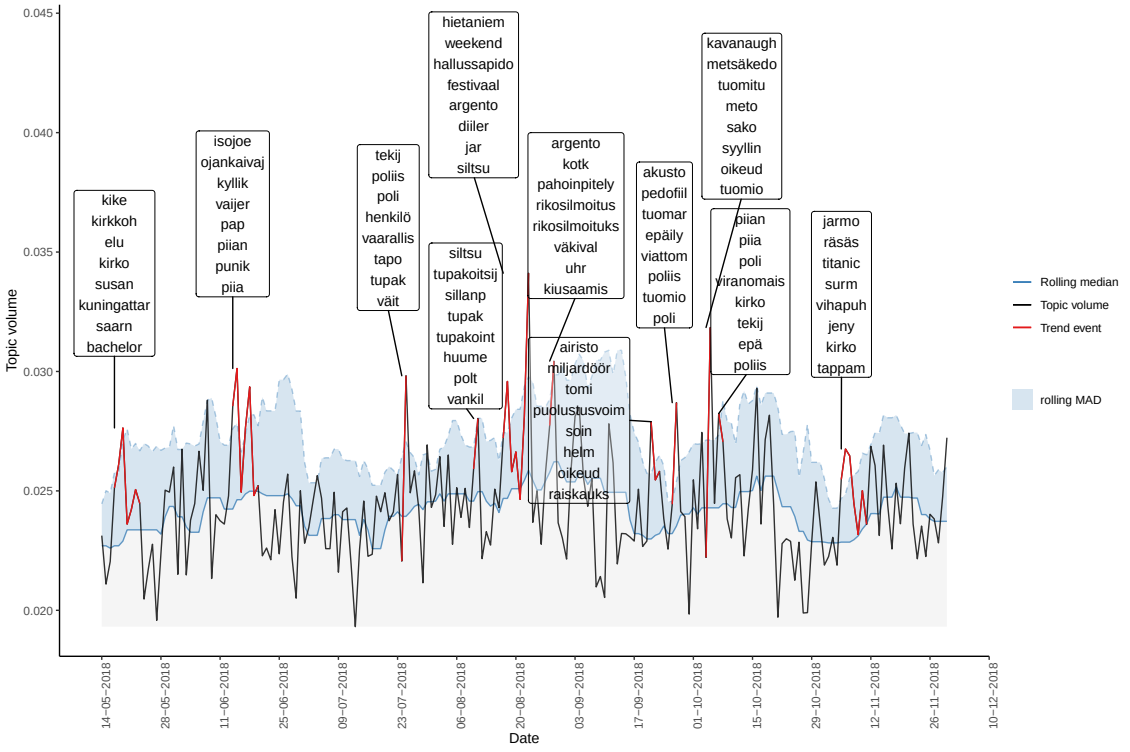


Figure B3: Vauva LDA 40, topic "Crime".

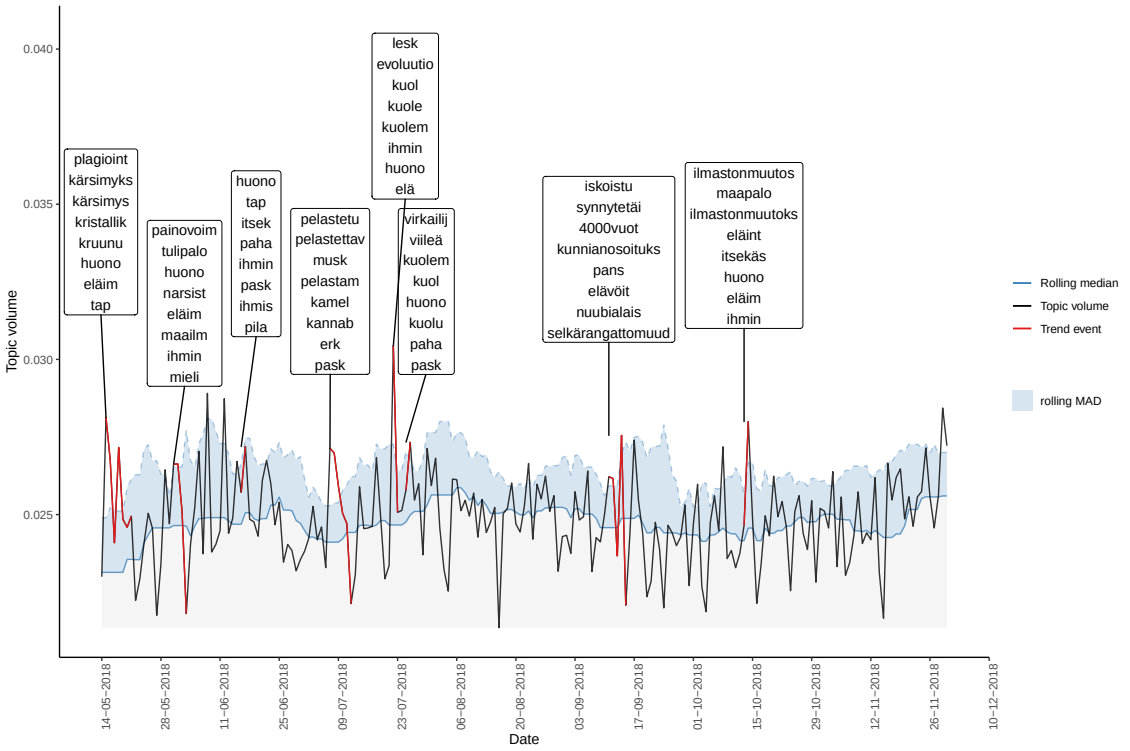


Figure B4: Vauva LDA 40, topic "Criticizing".

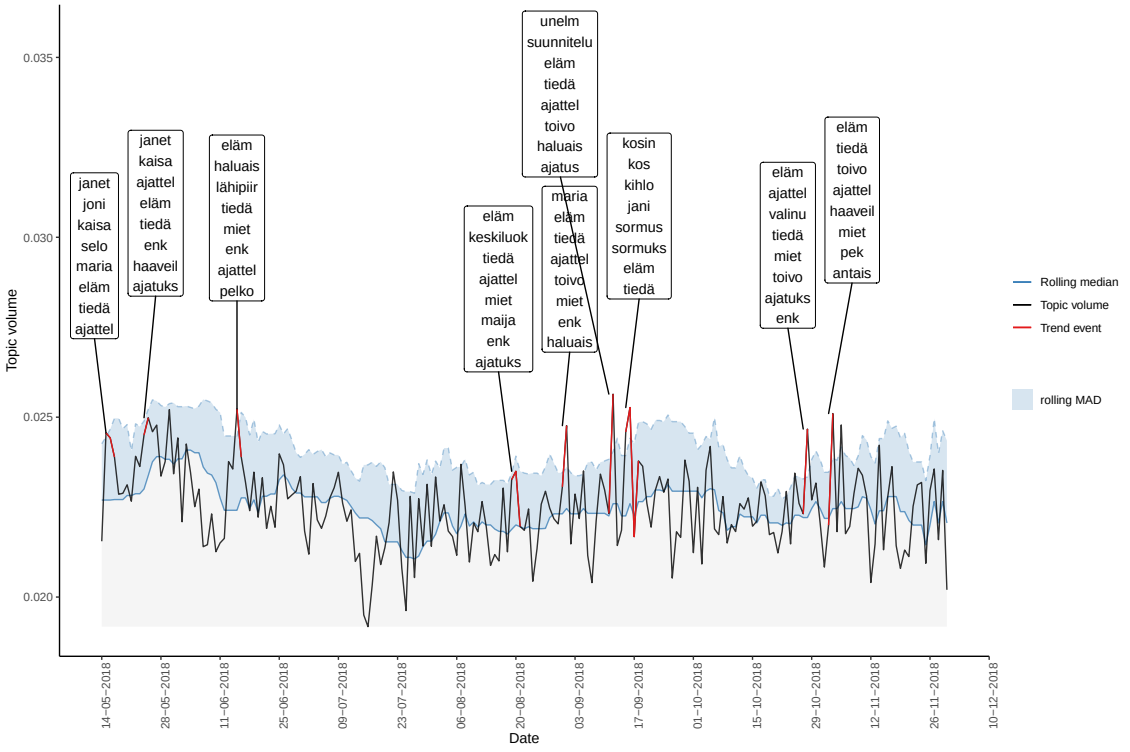


Figure B5: Vauva LDA 40, topic "Depression".

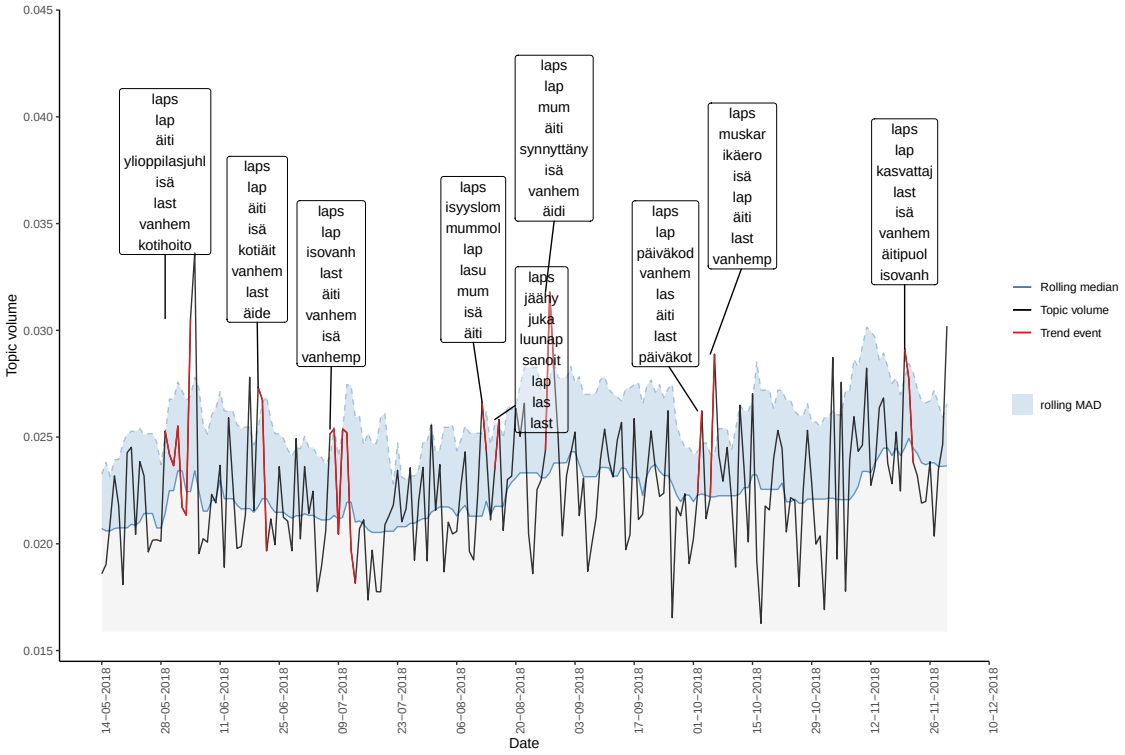


Figure B6: Vauva LDA 40, topic "Family".

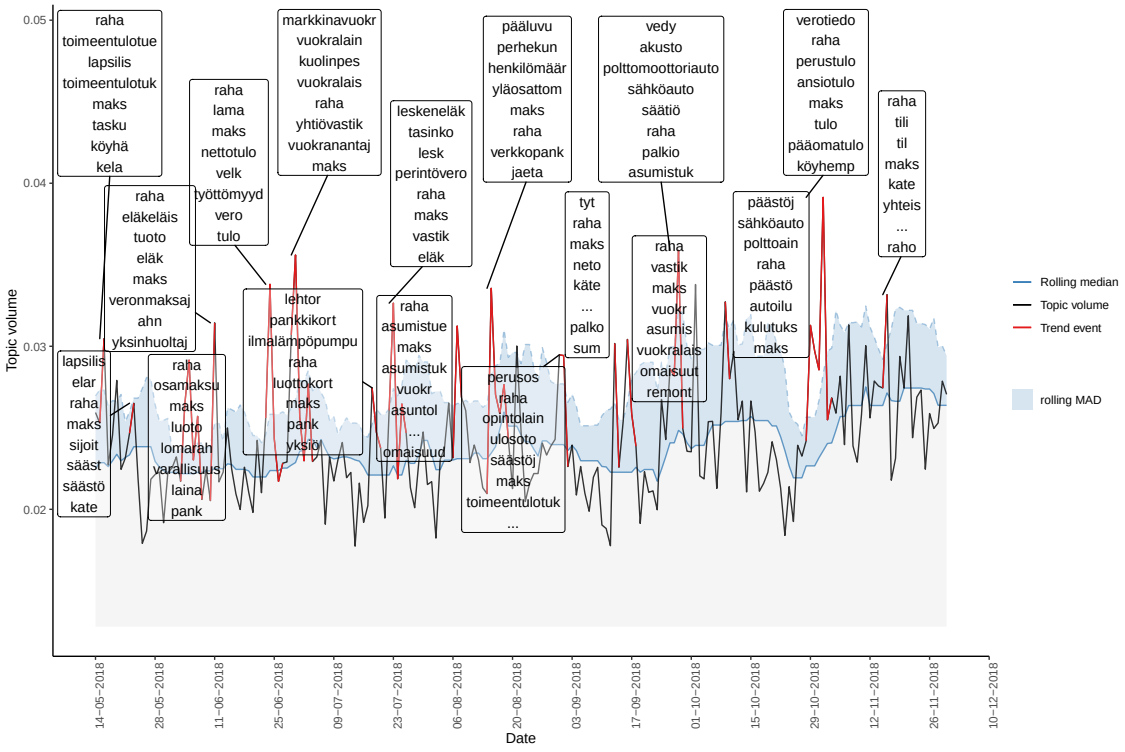


Figure B7: Vauva LDA 40, topic "Finances".

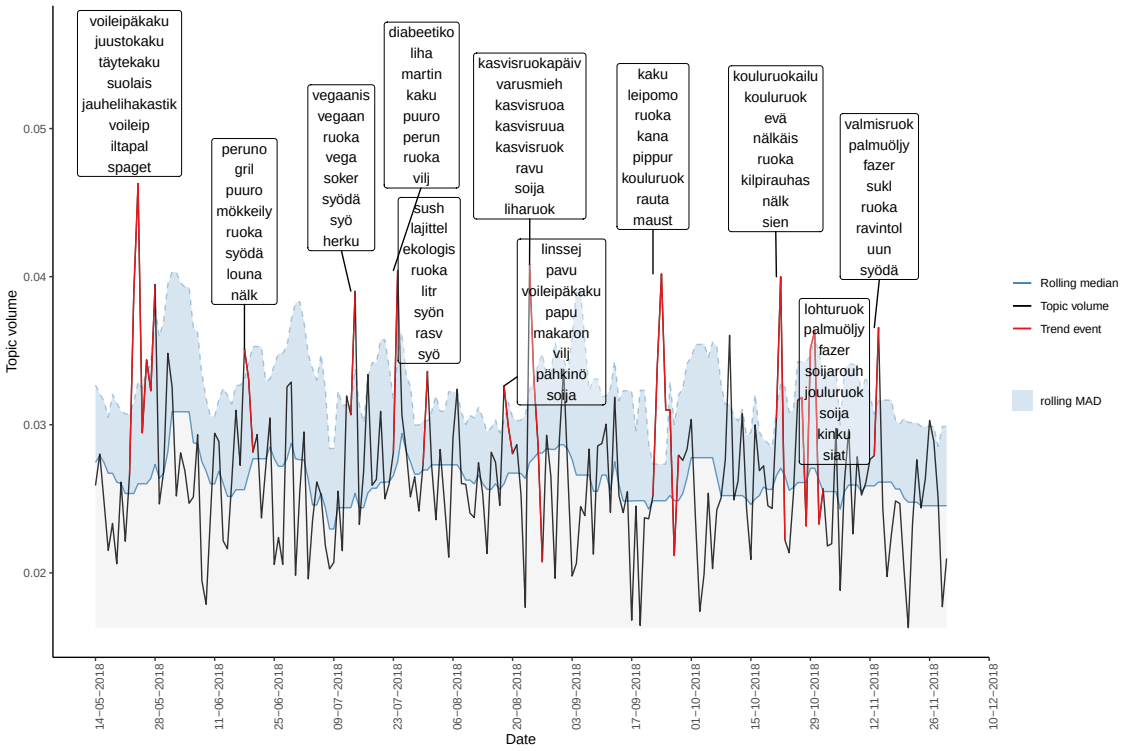


Figure B8: Vauva LDA 40, topic "Food".

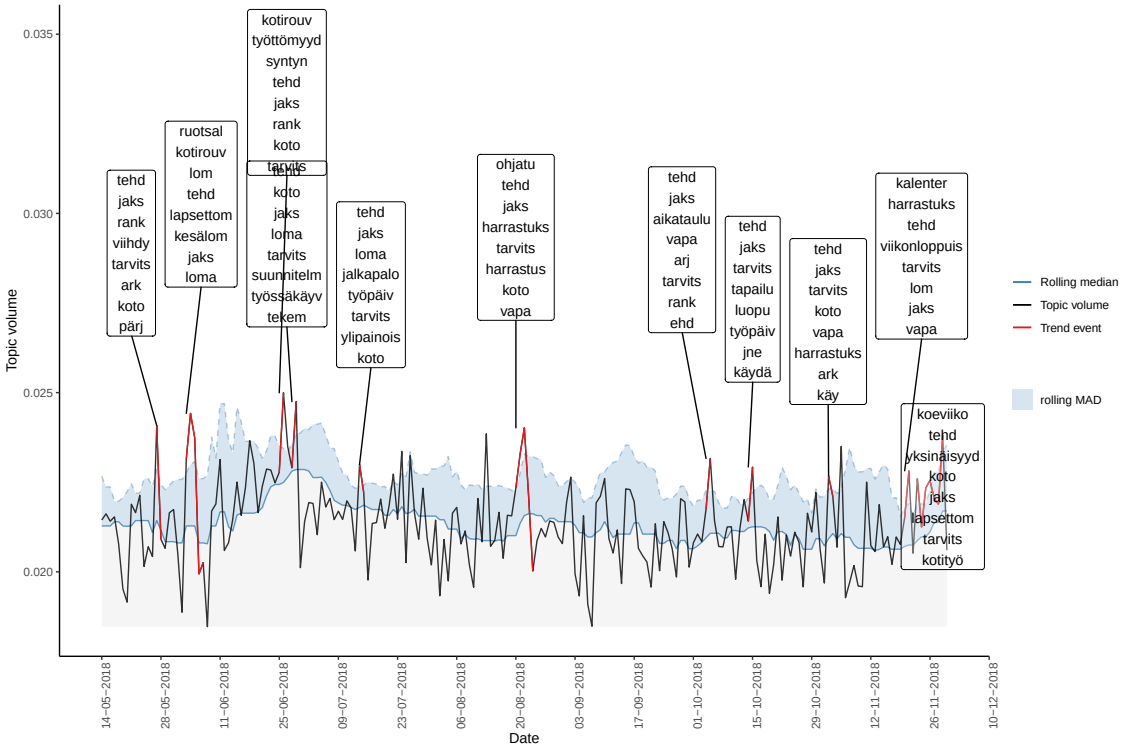


Figure B9: Vauva LDA 40, topic "Hobbies".

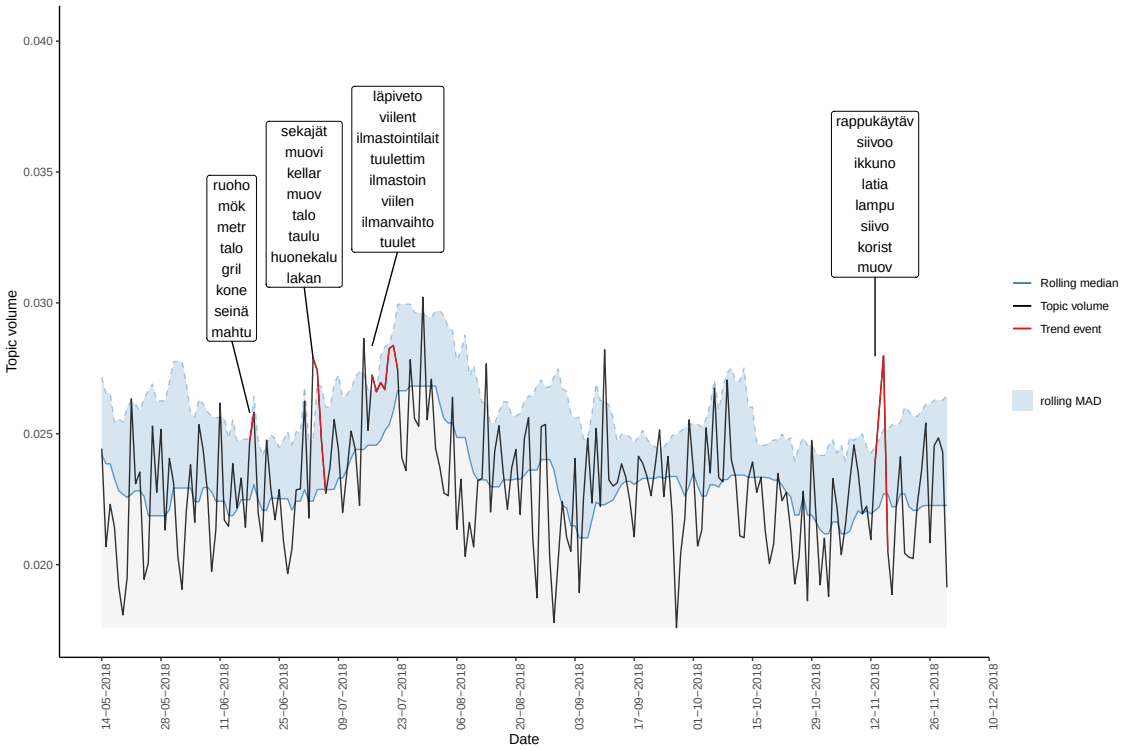


Figure B10: Vauva LDA 40, topic "Household".

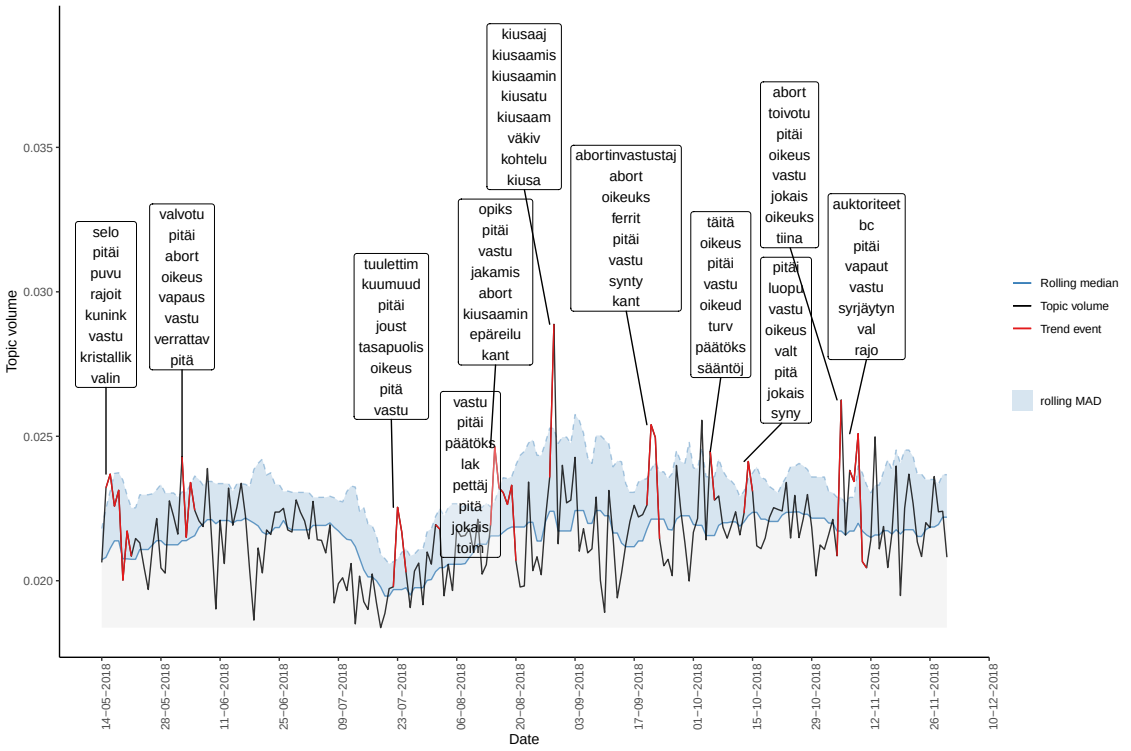


Figure B11: Vauva LDA 40, topic "Justice".

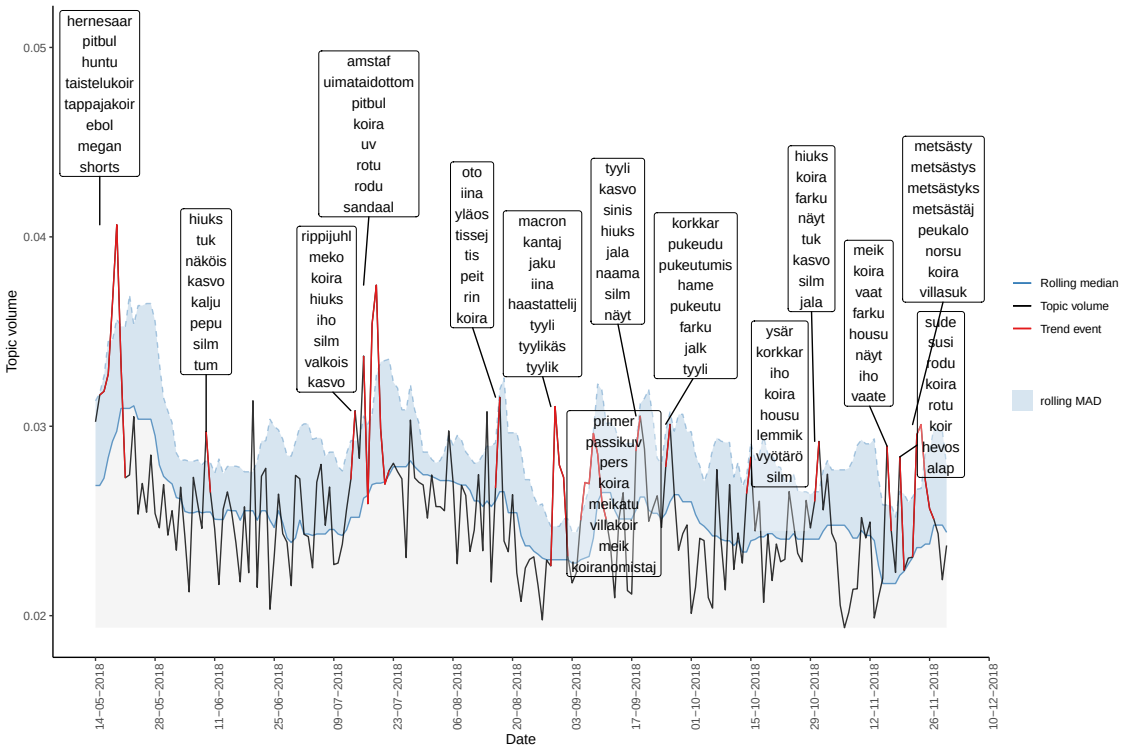


Figure B12: Vauva LDA 40, topic "Looks".

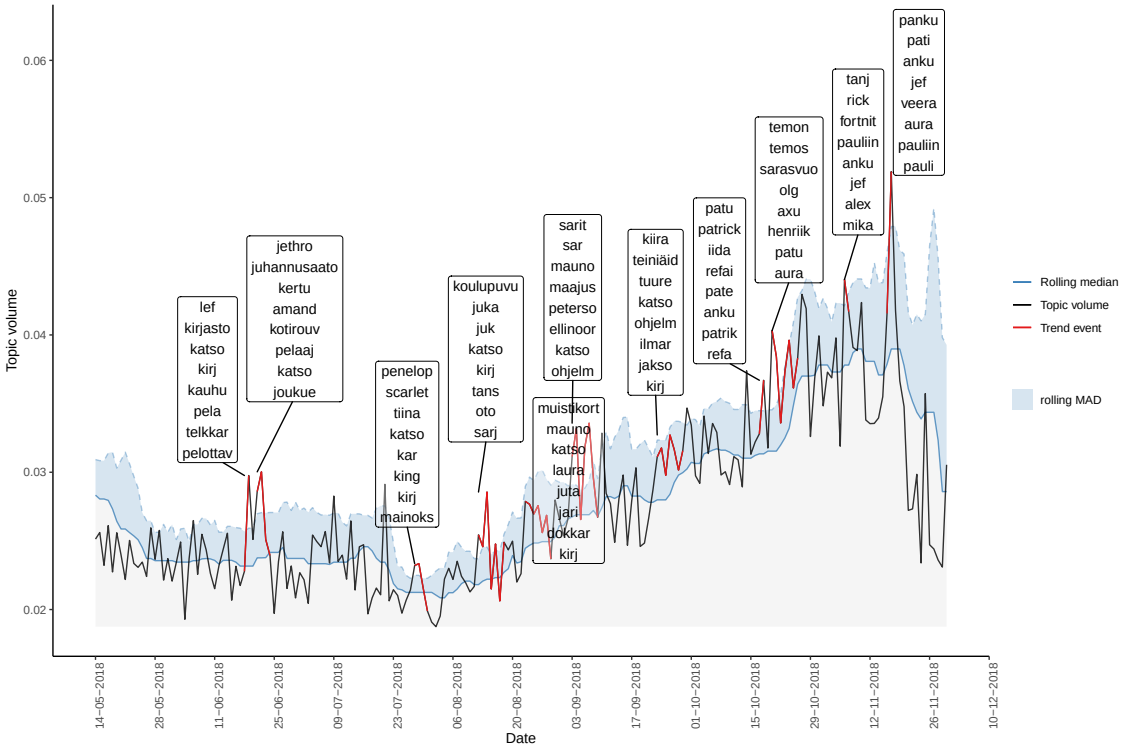


Figure B13: Vauva LDA 40, topic "Media".

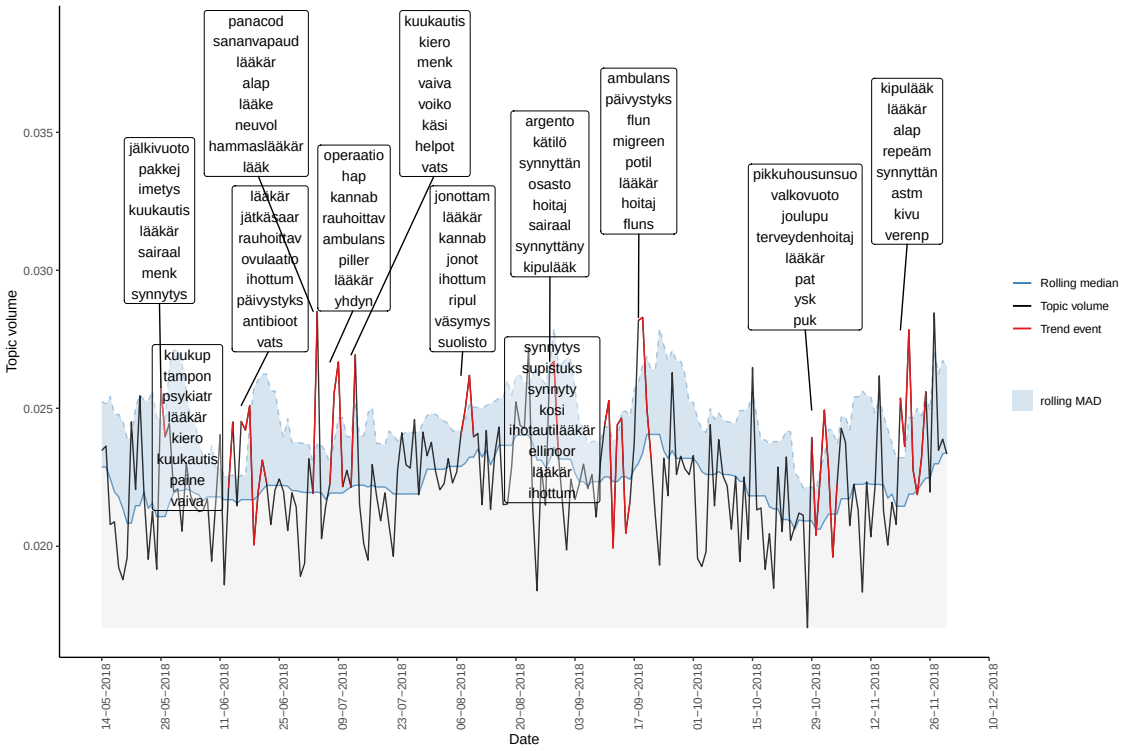


Figure B14: Vauva LDA 40, topic "Medical".

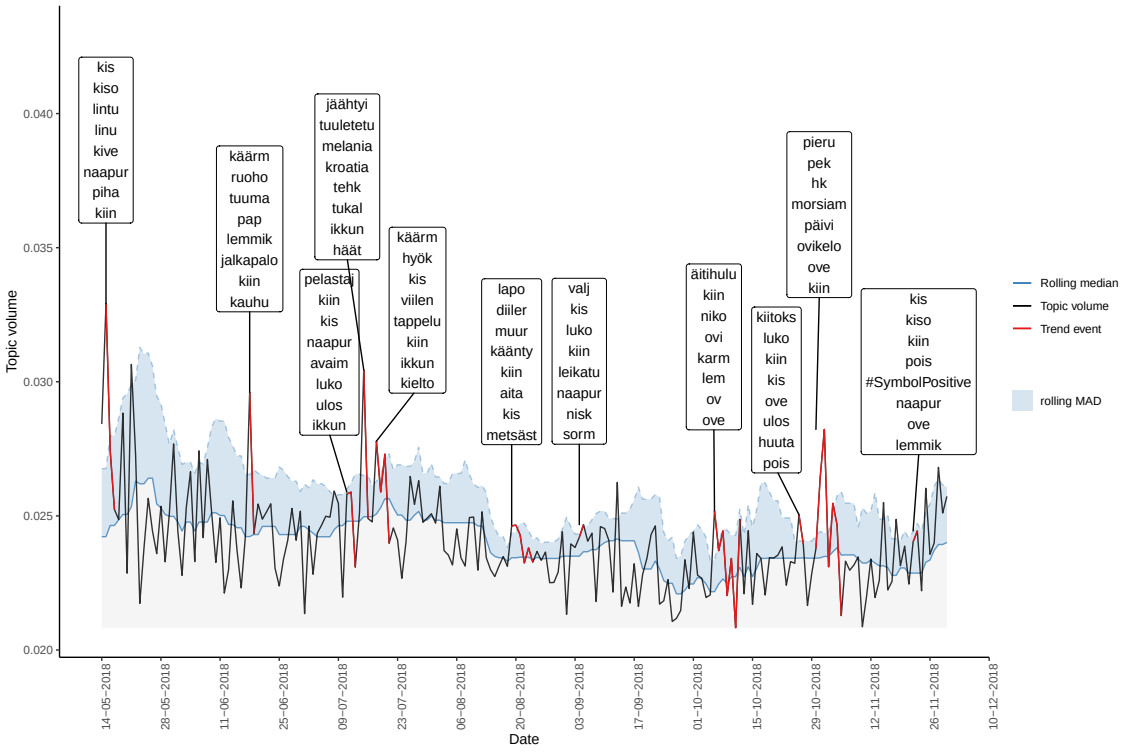


Figure B15: Vauva LDA 40, topic "Neighbors".

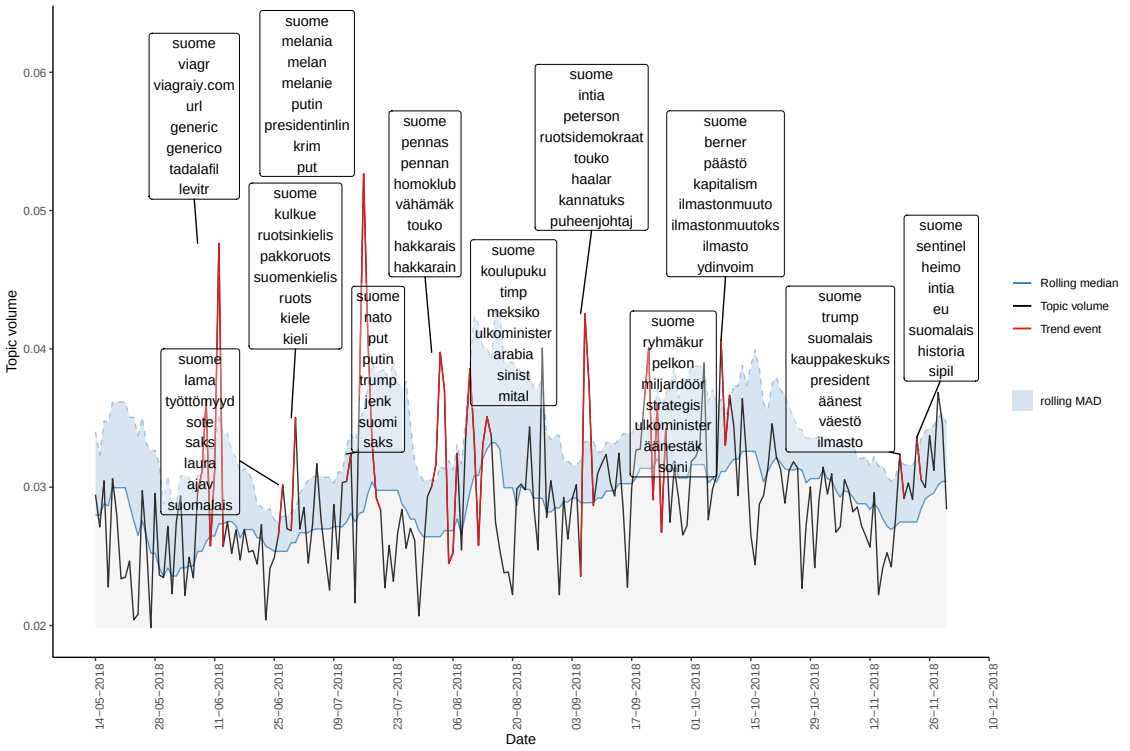


Figure B16: Vauva LDA 40, topic "Politics".

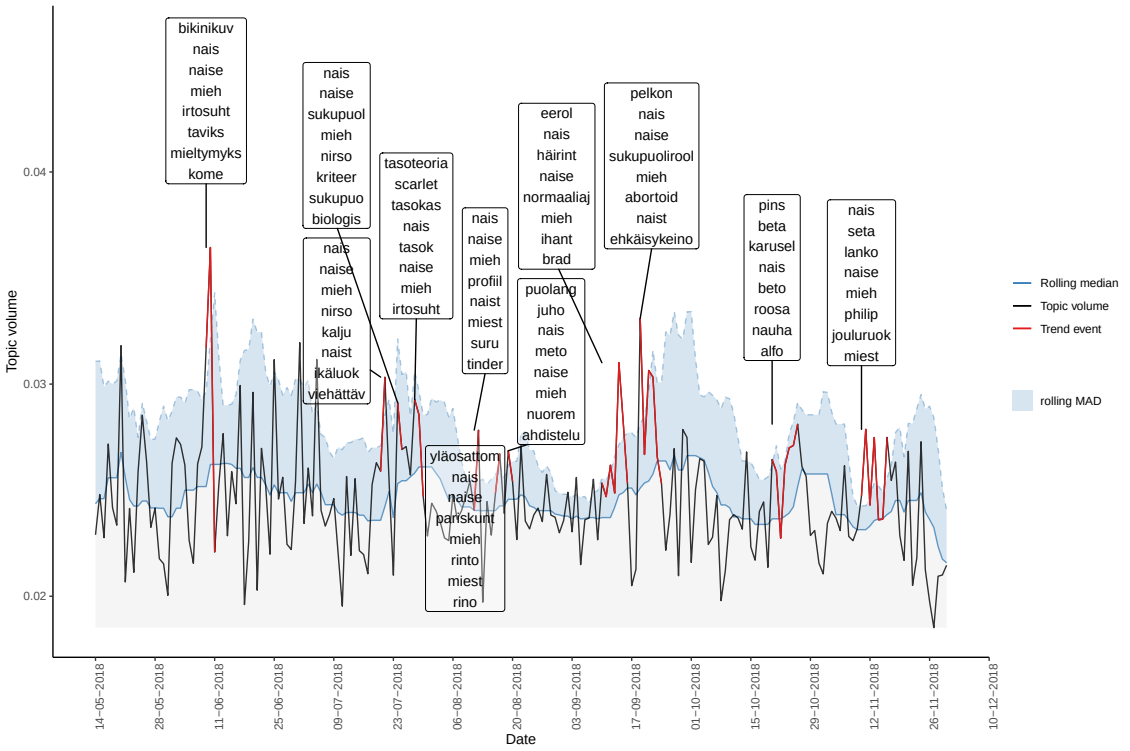


Figure B17: Vauva LDA 40, topic "Relationships".

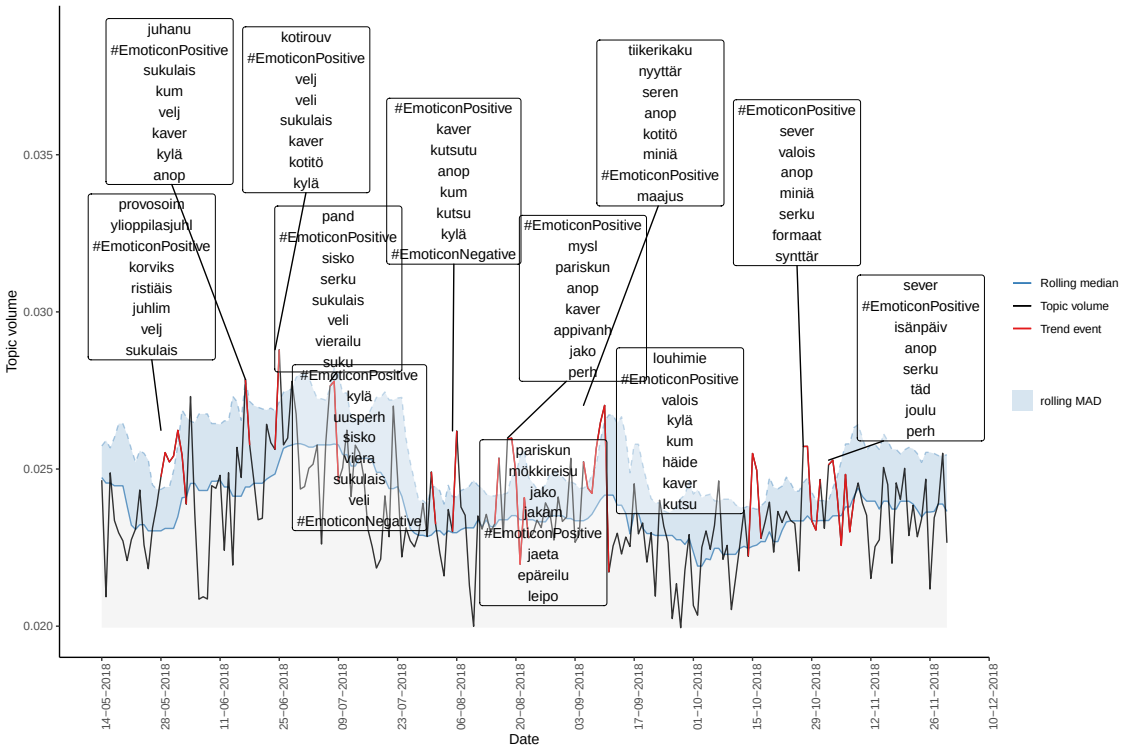


Figure B18: Vauva LDA 40, topic "Relatives".

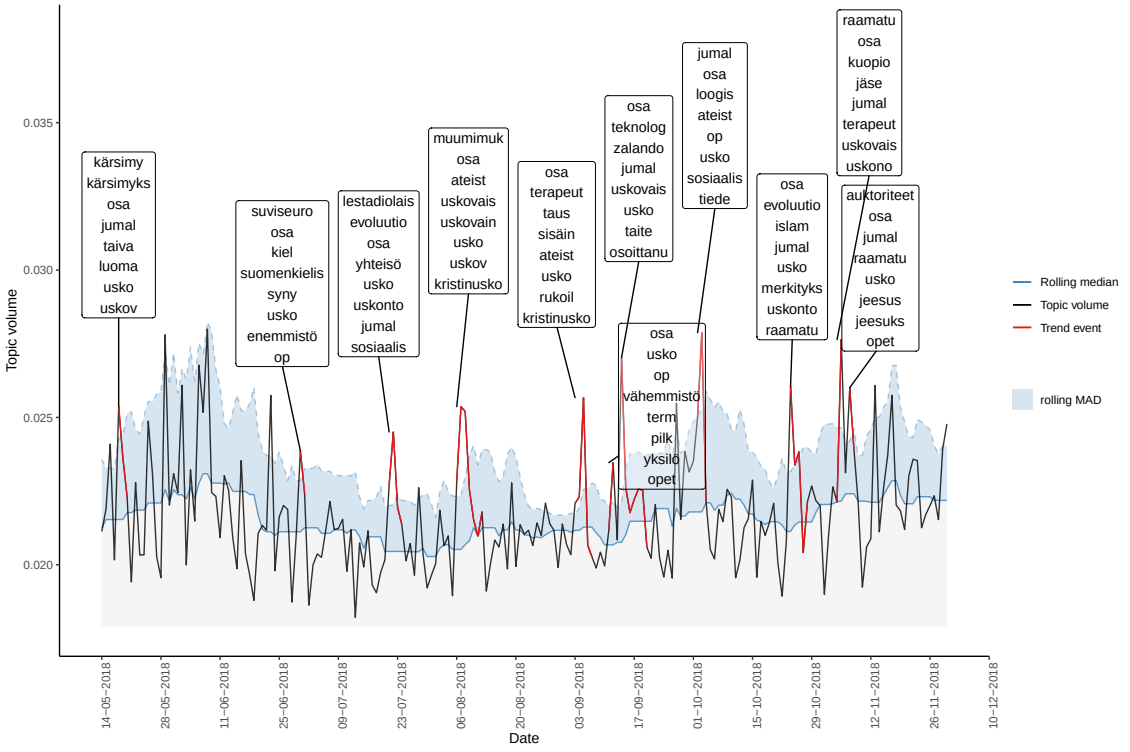


Figure B19: Vauva LDA 40, topic "Religion".

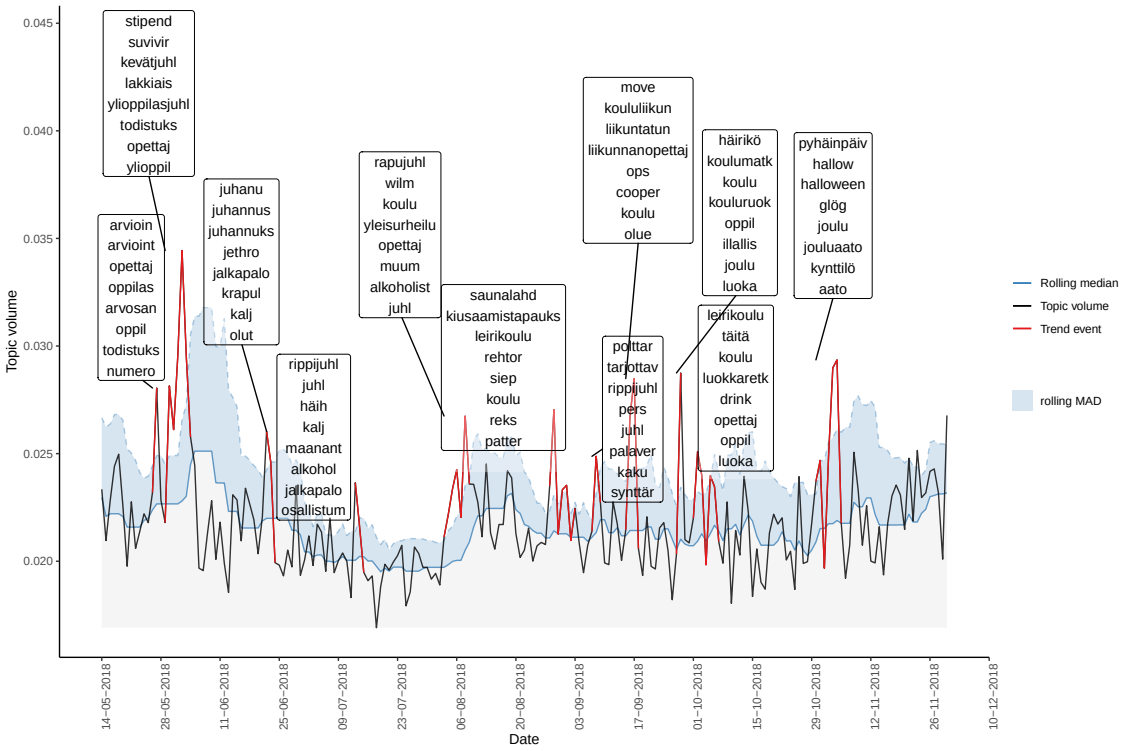


Figure B20: Vauva LDA 40, topic "Schooling".

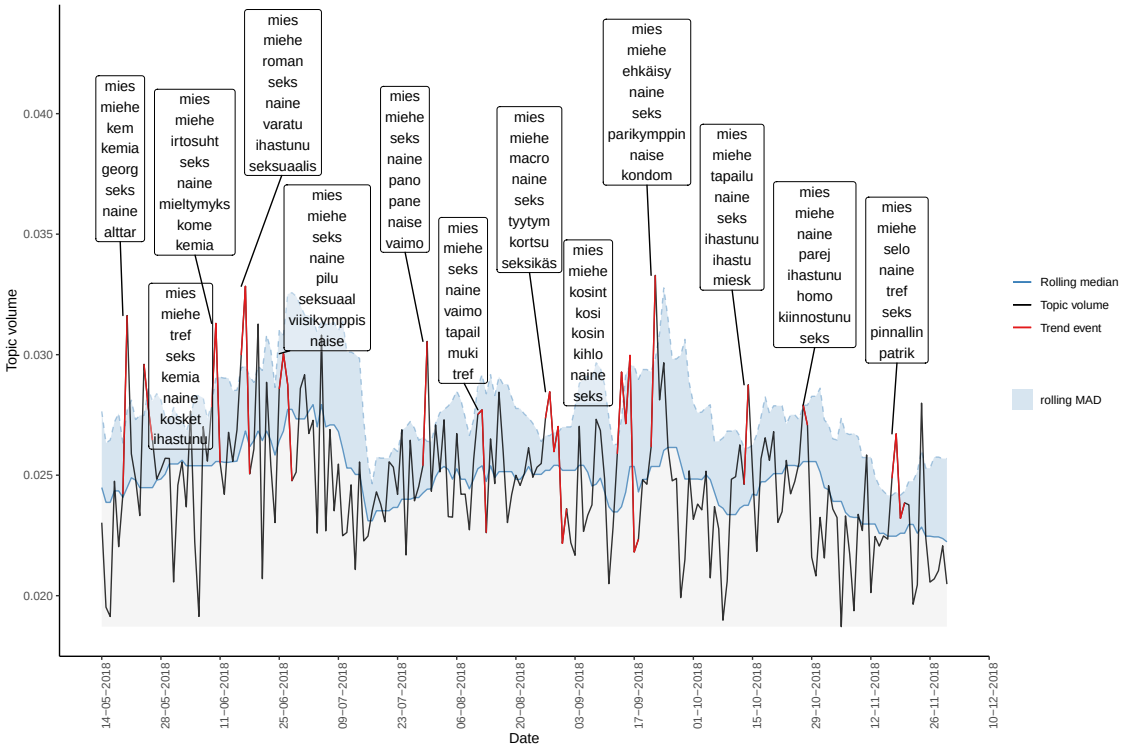


Figure B21: Vauva LDA 40, topic "Sex".

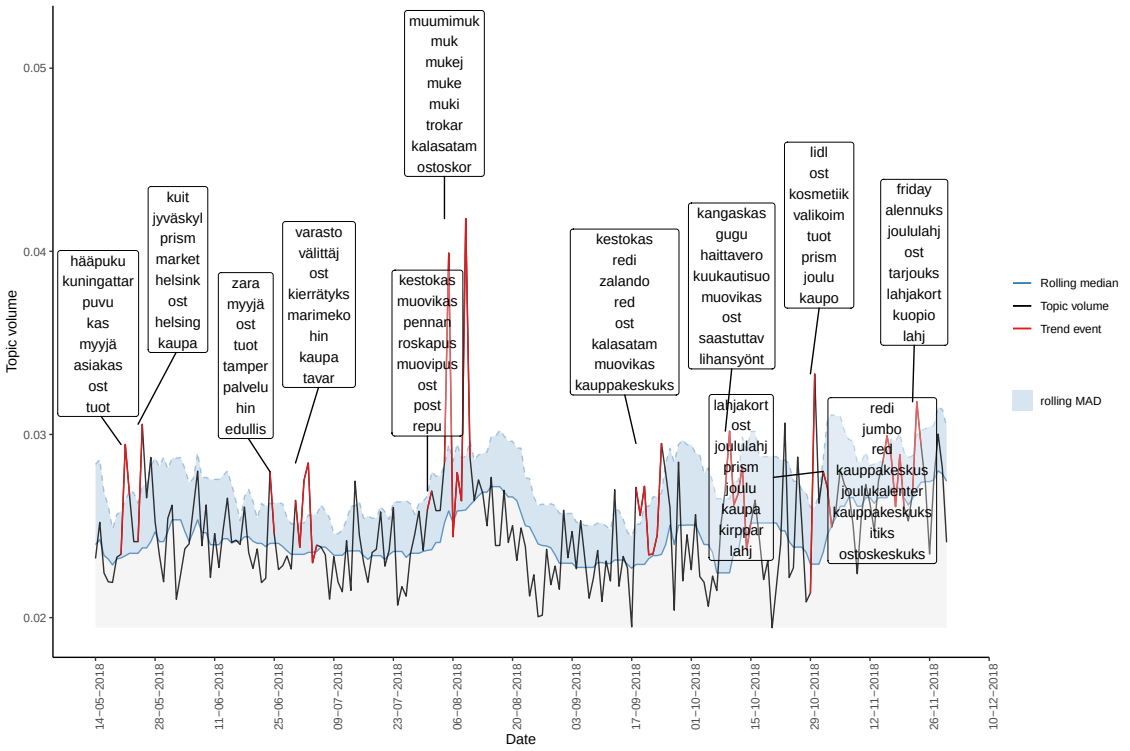


Figure B22: Vauva LDA 40, topic "Shopping".

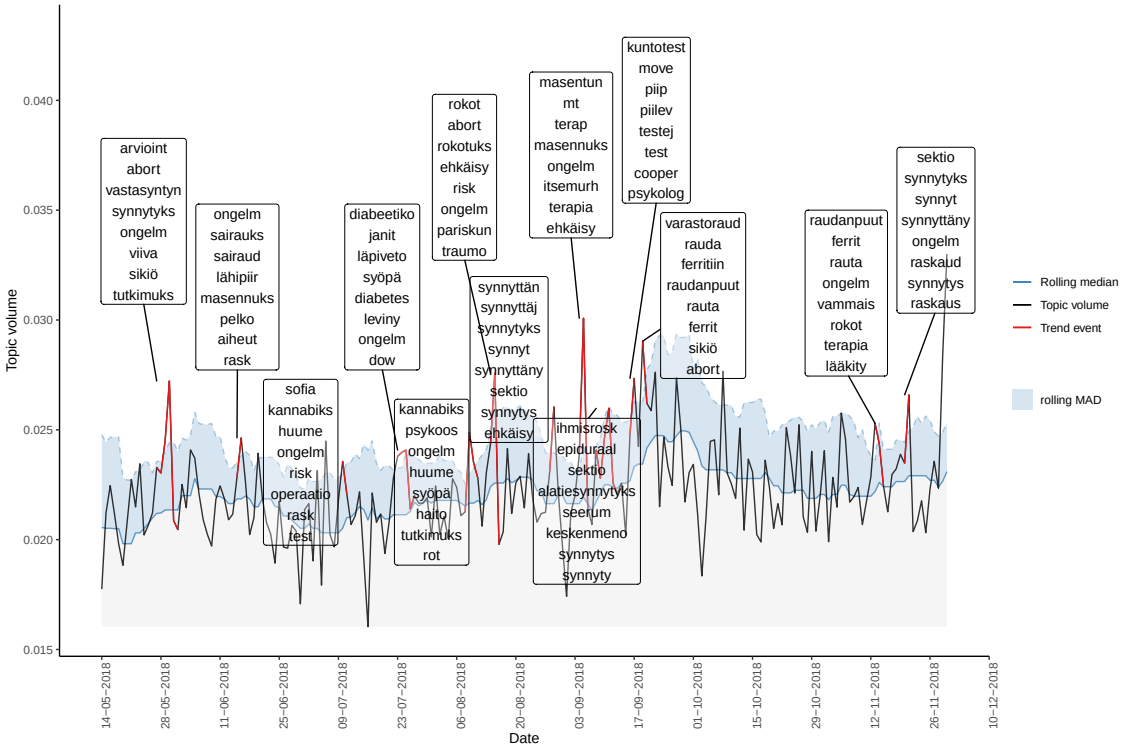


Figure B23: Vauva LDA 40, topic "Sickness".

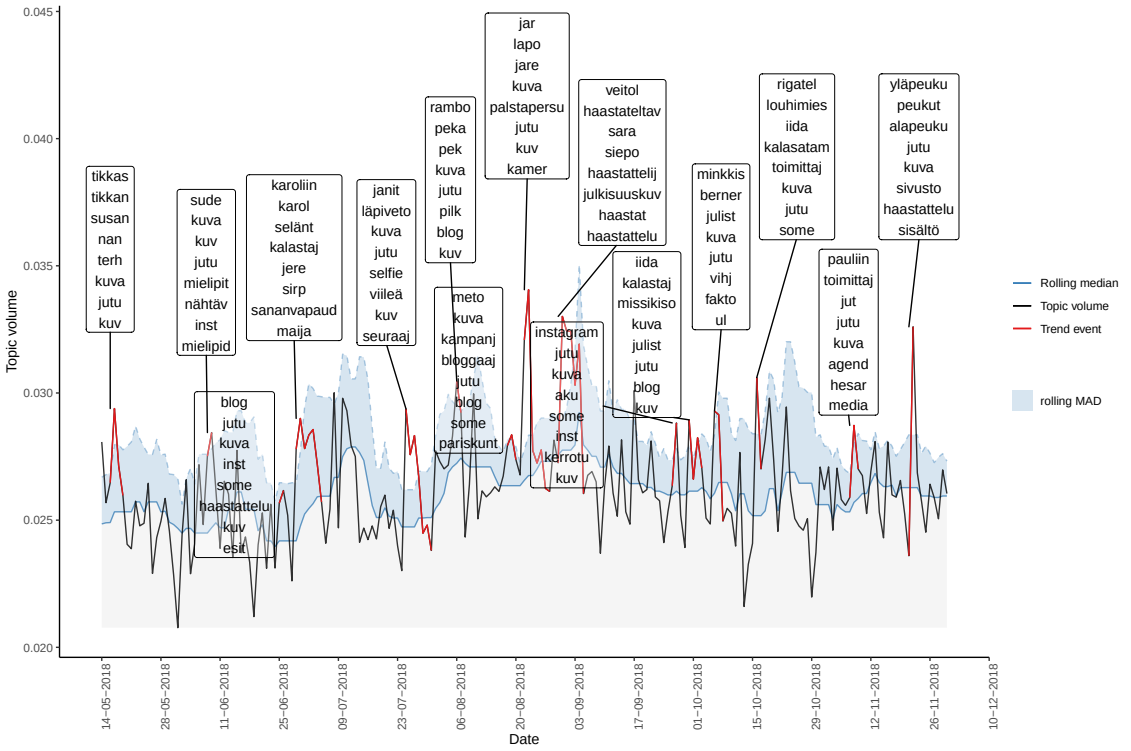


Figure B24: Vauva LDA 40, topic "Social media".

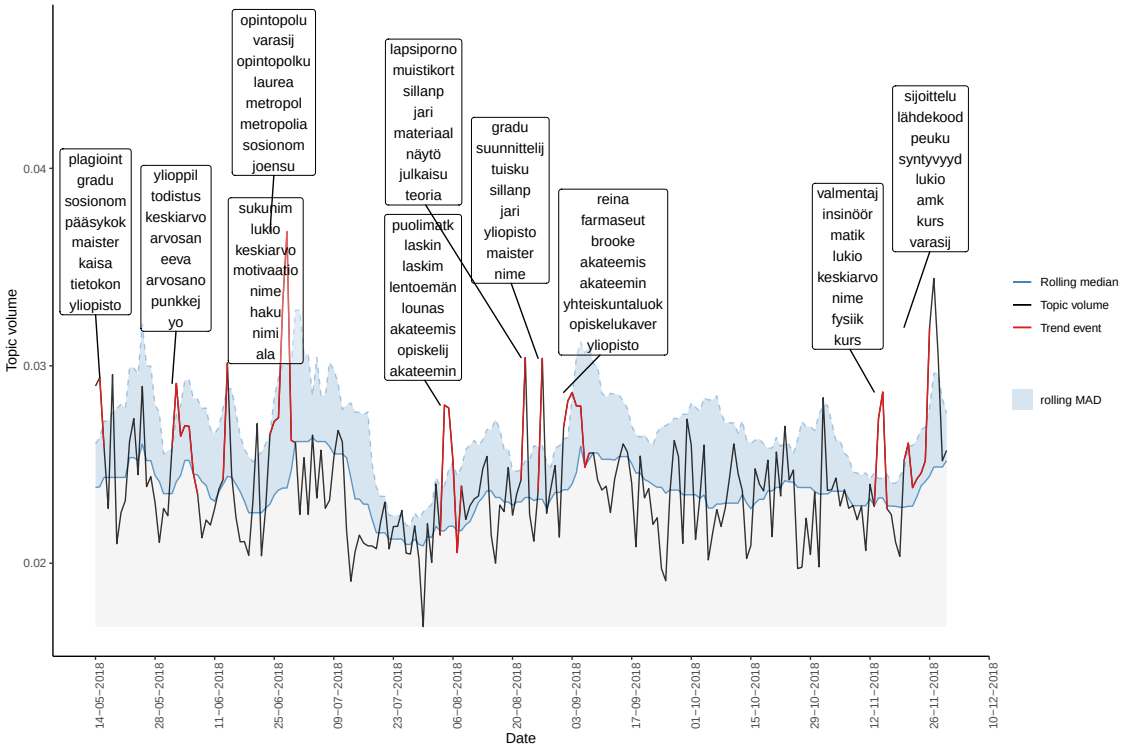


Figure B25: Vauva LDA 40, topic "Studying".

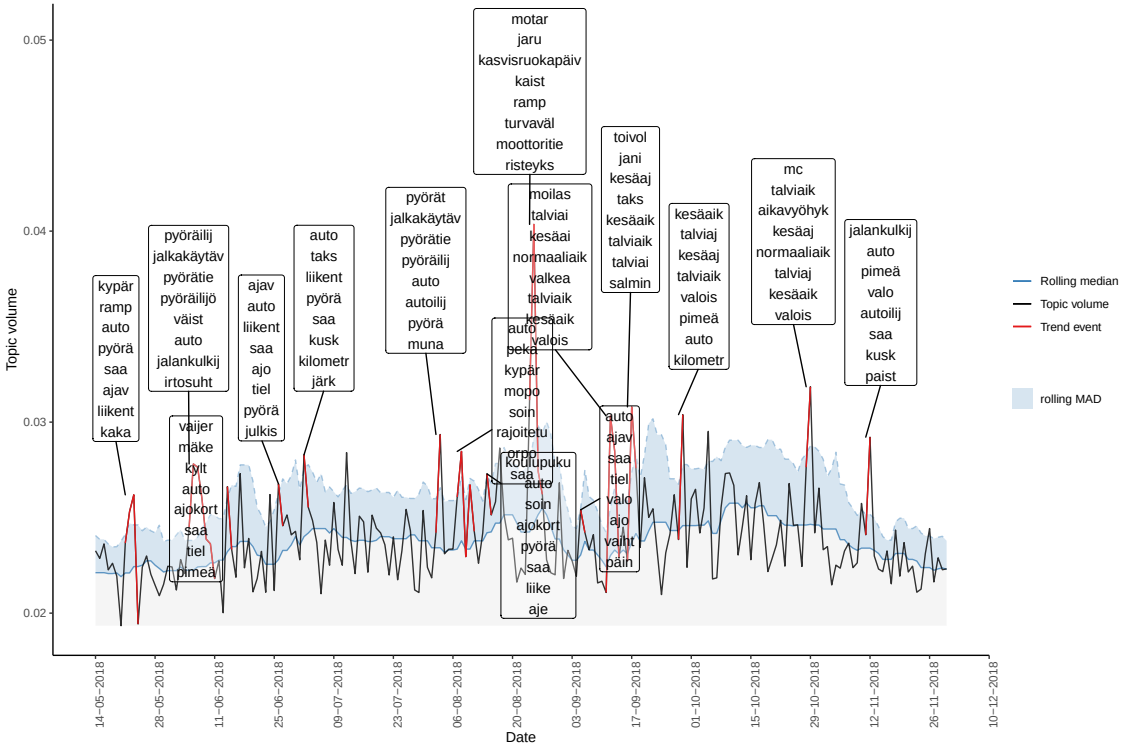


Figure B26: Vauva LDA 40, topic "Traffic".

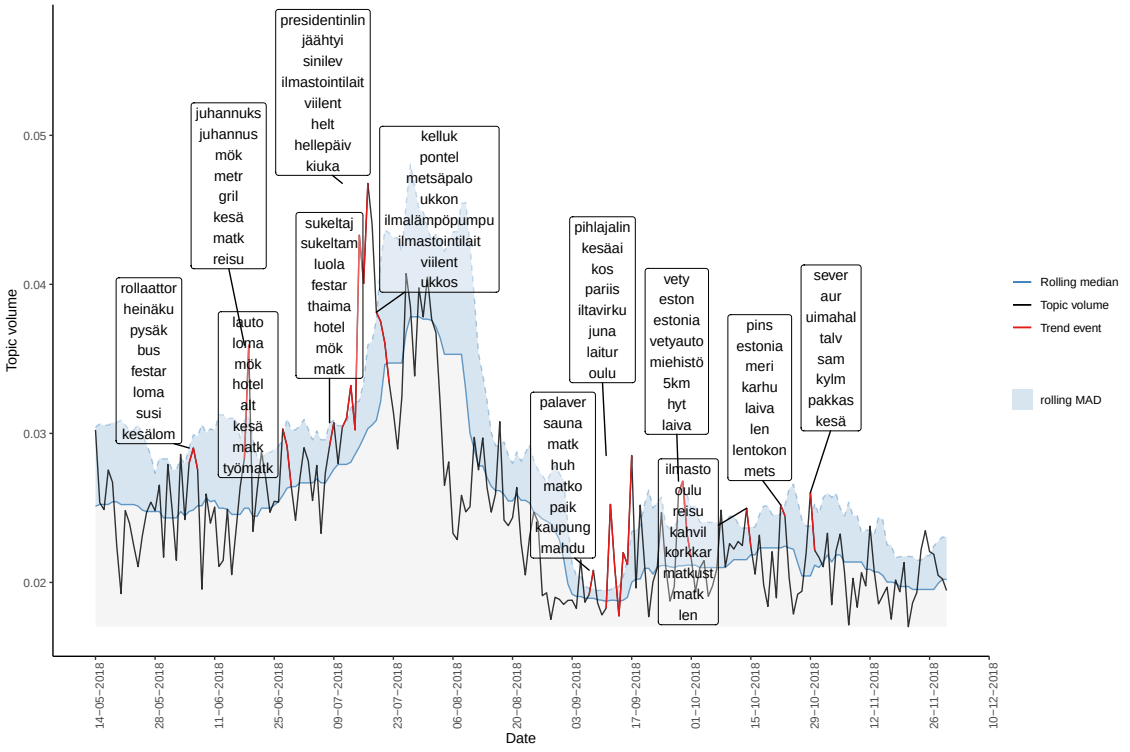


Figure B27: Vauva LDA 40, topic "Vacations".

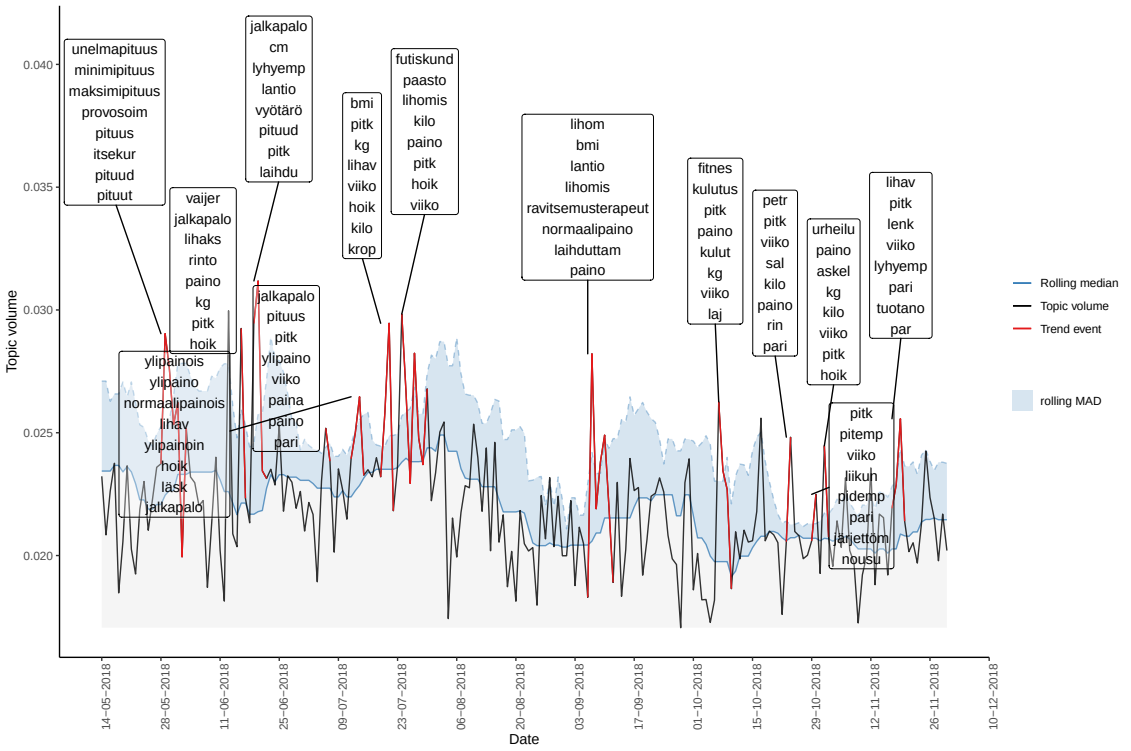


Figure B28: Vauva LDA 40, topic "Bodyweight".

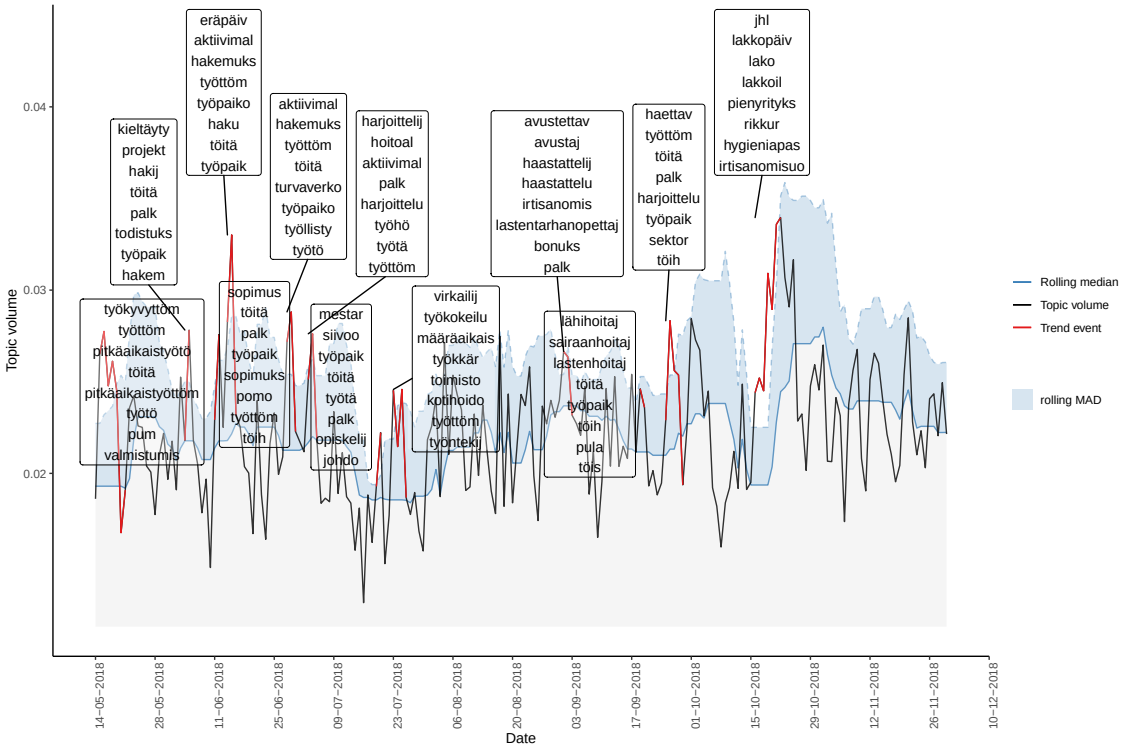


Figure B29: Vauva LDA 40, topic "Working life".